Slow and Drastic Change Detection in General HMMs Using Particle Filters with Unknown Change Parameters

Namrata Vaswani

Dept of Electrical and Computer Engineering and Center for Automation Research

University of Maryland, College Park, MD 20742, USA

namrata@cfar.umd.edu

Abstract

We study the change detection problem in general HMMs, when change parameters are unknown and the change could be gradual (slow) or sudden (drastic). Drastic changes can be detected easily using the increase in tracking error or the negative log of the observation likelihood conditioned on past observations (OL). But slow changes usually get missed. We propose a statistic for slow change detection called ELL which is the conditional Expectation of the negative Log Likelihood of the state given past observations. We show asymptotic stability (stability under weaker assumptions) of the errors in approximating the ELL for changed observations using a particle filter that is optimal for the unchanged system. It is shown that the upper bound on ELL error is an increasing function of the "rate of change" with increasing derivatives of all orders, and its implications are discussed. We also demonstrate, using the bounds on the errors, the complementary behavior of ELL and OL. Results are shown for simulated examples and for a real abnormal activity detection problem.

I. Introduction

Change or abnormality detection is required in many practical problems arising in quality control, flight control, fault detection and in surveillance problems like abnormal activity detection [1], [2]. In most cases, the underlying system in its normal state can be modeled as a parametric stochastic model. The observations are usually noisy (making the system partially observed). Such a system forms a "general HMM" [3] (also referred to as a "partially observed nonlinear dynamical model" or a "stochastic state space model" in different contexts). It can be approximately tracked (estimate probability distribution of hidden state variables given observations) using a Particle Filter (PF) [4].

The author would like to acknowledge Professors Chellappa, Papamarcou and Slud for their valuable suggestions.

We study the change detection problem in a general HMM when the change parameters are unknown and the change can be slow or drastic. We use a PF to estimate the posterior probability distribution of the state at time t, X_t, given observations up to t, $Pr(X_t \in dx|Y_{1:t}) \stackrel{\triangle}{=} \pi_t(dx)$. Drastic changes can be detected easily using the increase in tracking error or the negative log of observation likelihood (OL). But slow changes usually get missed. We propose a statistic called ELL (Expected Log-Likelihood) which is able to detect slow changes. ELL is the conditional Expectation of the negative Log-Likelihood of the state at time t ($[-\log p_t^0(X_t)]$), given past observations, $Y_{1:t}$ i.e. it is the expectation under π_t of $[-\log p_t^0(X_t)]$.

We show in Section III-A that ELL is equivalent to the Kerridge Inaccuracy [5] between the posterior and prior state distributions. Averaging the log likelihood over a time sequence of i.i.d. observations is often used in hypothesis testing and in [6] it is shown to be equivalent to the Kerridge Inaccuracy between the empirical distribution of the i.i.d. observations and their actual pdf. But to the best of our knowledge, ELL defined as the expectation of log likelihood of state given past observations, in the context of HMMs (and its estimation using a PF) has not been used before.

Now, ELL detects a slow change before the PF loses track. This can be useful in any target(s) tracking problem where the target(s) dynamics might change over time. If one can detect the change, one can learn its parameters on the fly and use the changed system model (or at least increase the system noise variance to track the change), without losing track of the target(s). We have used ELL to detect changes in landmark shape dynamical models [1], [2] and this has applications in abnormal activity detection, medical image processing (detecting motion disorders by tracking patients' body parts) and activity segmentation (segmenting a long activity sequence into piecewise stationary elementary activities) [2]. We briefly discuss the abnormal activity detection problem in Section VIII-B. Other applications of ELL which we are working on currently, are in neural signal processing (detecting changes in response of animals' brains to changes in stimuli provided to them). ELL can also potentially be used for congestion detection since congestion quite often starts as a slow change.

A. The General HMM Model

We assume a general HMM [3] with an $\Re^{n_x}$ valued state process $X = \{X_t\}$ and an $\Re^{n_y}$ valued observation process $Y = \{Y_t\}$[1]. The system (or state) process $\{X_t\}$ is a Markov process with state transition kernel $Q_t(x_t, dx_{t+1})$ and the observation process is a memoryless function of the state given by $Y_t = h_t(X_t) + w_t$

[1]We use the subscript 't' (e.g. X_t, Y_t) instead of 'n' for (discrete) time instants, to avoid confusion with N used for the number of particles in Particle Filtering

where w_t is an i.i.d. noise process and h_t is, in general, a nonlinear function. The state dynamics defined by Q_t can also be linear or nonlinear. We denote the conditional distribution of the observation given state by $G_t(dy_t, x_t)$. It is assumed to be absolutely continuous [7] and its pdf is given by $g_t(Y_t, x) \triangleq \psi_t(x)$. The prior initial state distribution, $p_0(x)$, the conditional distribution of observation given state and the state transition kernel are known and assumed to be absolutely continuous[2]. Thus the prior distribution of the state at any t is also absolutely continuous and admits a density, $p_t(x)$.

B. Problem Definition

We study the problem of detecting slow and drastic changes in the system model of a general HMM described above, when the change parameters are unknown. We assume that the normal (original/unchanged) system has state transition kernel Q_t^0. A change in the system model begins to occur at some finite time t_c and lasts till a final finite time t_f. In the time interval, $[t_c, t_f]$, the state transition kernel is Q_t^c and after t_f it again becomes Q_t^0. Both Q_t^c and the change start and end times t_c, t_f are assumed to be unknown. The goal is to detect the change, with minimum delay. Note that although the change in system model lasts for a finite time, $[t_c, t_f]$, its effect on the prior state pdf $p_t^0(x)$ is either permanent or it lasts for a much longer time. very slowly mixing).

C. Related Work

For linear dynamical systems with known changed system parameters, the CUSUM (cumulative sum) [8] algorithm can be used directly. The CUSUM algorithm uses as change detection statistic, the maximum (taken over all previous time instants) of the likelihood ratio assuming that the change occurred at time j, i.e. $CUSUM_t \triangleq \max_{1\leq j\leq t} LR(j),\ LR(j) = \frac{p_{\theta_1}(Y_j, Y_{j+1} \ldots Y_t)}{p_{\theta_0}(Y_j, Y_{j+1} \ldots Y_t)}$. For unknown changed system parameters, the Generalized Likelihood Ratio Test can be used whose solution for linear systems in well known [8]. When a nonlinear system experiences a change, linearization techniques like extended Kalman filters and change detection methods for linear systems are the main tools [8]. Linearization techniques are computationally efficient but are not always applicable.

In [9], the authors attempt to use a particle filtering approach for sudden change detection in nonlinear systems without linearization. They define a modification of the CUSUM change detection statistic that can be efficiently evaluated using PFs. Both CUSUM and the statistic of [9] assume known change parameters and are based on the likelihood ratio of the current (t-j+1) observations, $LR(j)$. An entirely different class of approaches (e.g. see [10]) used extensively with PFs uses a discrete state variable to denote the mode that the system is operating in. When

[2]Note that for ease of notation, we denote the pdf either by the same symbol or by the lowercase of the probability distribution symbol

changed system parameters are not known, sudden changes can be detected using tracking error [11] which is the distance (usually Euclidean distance) between the current observation and its prediction based on past observations. These and some other approaches for sudden change detection using PFs are discussed in a recent survey article [12].

In this work, we have also studied the stability of errors in approximating the ELL for changed observations using a PF that is optimal for the unchanged system. There has been a lot of recent research on studying the stability of the optimal nonlinear filter. Asymptotic stability results w.r.t. initial condition were first proved in [13]. The Hilbert projective metric has been used to prove stability w.r.t. the initial condition and also w.r.t. the model [14], [15]. New approaches have been proposed recently for noncompact state spaces [16], [17]. The results for stability w.r.t. the model have been used to prove convergence of the PF estimate of the posterior with number of particles, $N \to \infty$ [3], [18]. We use in this work, results from [3] in which the authors have replaced the mixing transition kernel assumption required for proving stability with a much weaker mixing unnormalized filter kernel assumption.

D. Organization of the Paper

We discuss some notation, definitions and the particle filtering algorithm in Section II. ELL, its relation with Kerridge Inaccuracy, the use of the OL statistic for cases where ELL fails and certain practical issues are discussed in Section III. In Section IV, we show asymptotic stability and stability (under weaker assumptions) of the errors in ELL approximation using a PF optimal for the unchanged system. In Section V, we bound the ELL approximation error by an increasing function of the rate of change and discuss its implications. We discuss complementary behavior of ELL and OL for slow and drastic changes in Section VI. A simple example is analyzed and generalizations of the theorems in this paper are discussed in Section VII. We present simulation results and results for abnormal activity detection [1], [2] in Section VIII and give conclusions in Section IX.

II. NOTATION AND PRELIMINARIES

A. Notation and Definitions

We use H_0 to denote the original or unchanged system hypothesis and H_c to denote the changed system hypothesis. Also, the superscript ^{c} is used to denote any parameter related to the changed system, ^{0} for the original system and ^{c,0} for the case when the observations of the changed system are filtered using a filter optimal for the original system[3]. Thus the posteriors, $\pi_t^{0,0}(dx) = Pr(X_t \in dx | Y_{1:t}^0, H_0)$ (also denoted by

[3]At most places ^{0,0} is replaced by ^{0} and ^{c,c} by ^{c}

π_t^0), $\pi_t^{c,c}(dx) = Pr(X_t \in dx|Y_{1:t}^c, H_c)$ (also denoted by π_t^c) and $\pi_t^{c,0}(dx) = Pr(X_t \in dx|Y_{1:t}^c, H_0)$ where $Y_{1:t}^c = (Y_{1:t_c-1}^0, Y_{t_c:t}^c)$ $\forall t \le t_f$ and $Y_{1:t}^c = (Y_{1:t_c-1}^0, Y_{t_c:t_f}^c, Y_{t_f+1:t}^0)$ $\forall t > t_f$. Also, for PF estimates of these distributions, we add a superscript ^{N} to denote number of particles, for e.g. $\pi_t^{0,N}$, $\pi_t^{c,N}$, $\pi_t^{c,0,N}$.

With any nonnegative kernel, J, defined on the state space E, is associated a nonnegative linear operator denoted by J and defined by $J(\mu)(dx') \stackrel{\triangle}{=} \int_E \mu(dx)J(x,dx')$ for any nonnegative measure μ [3]. For any finite measure, μ, the normalized measure is denoted by $\bar{\mu} \stackrel{\triangle}{=} \mu/\mu(E)$. The normalized nonnegative nonlinear operator $\bar{J}$ is defined by $\bar{J}(\mu) \stackrel{\triangle}{=} \frac{J(\mu)}{J(\mu)(E)}$ [3]. Also, $(.,.)$ is the inner product notation.

The prior state distribution at t, $(Q_t^0(...(Q_1^0(\pi_0))))(dx)$ has pdf $p_t^0(x)$ while the changed system's prior state distribution, $(Q_t^0...(Q_{t_f}^c...(Q_{t_c}^c...(Q_1^0(\pi_0)))))(dx)$ has pdf $p_t^c(x)$. We discuss this in Section III-D.

Note that **"event occurs a.s."** refers to the event occurring almost surely w.r.t. the measure corresponding to the probability distribution of $Y_{1:t}$. Also, E_μ denotes expectation under the measure μ, for example E_{π_t} is expectation under the posterior state distribution. E_Y denotes expectation under the distribution of the random variable Y, for example $E_{Y_{1:t}}$ denotes expectation under the distribution of the observation sequences. Finally, Ξ_{pf} denotes averaging over different realizations of the PF each of which produces a different realization of the random measure π_t^N (expectation under the probability distribution of the random measure π_t^N).

Also note that we refer to Theorem x, part y as Theorem x.y (e.g. Theorem 1.1). We now present some definitions of terms used in the paper:

Definition 1: The **unnormalized filter kernel** [3] for a system with state transition kernel Q_t and probability of observation given state ψ_t, is given by $R_t(x,dx') = Q_t(x,dx')\psi_t(x')$. So $R_t^0 = Q_t^0\psi_t^0$ is the unnormalized filter kernel for the original system observations estimated using the original system model, Q_t^0; $R_t^c = Q_t^c\psi_t^c$ is the unnormalized filter kernel for the changed system observations using the changed system model, Q_t^c; while $R_t^{c,0} = Q_t^0\psi_t^c$ is the unnormalized filter kernel for the changed system observations using the original system transition kernel, Q_t^0.

Definition 2: A nonnegative kernel J defined on E is **mixing** [3] if there exists a constant, $0 < \epsilon < 1$ and a nonnegative measure λ s.t. $\epsilon\lambda(A) \le J(x,A) \le \frac{1}{\epsilon}\lambda(A)$ $\forall x \in E$ and for any Borel subset $A \subset E$. A sequence of mixing kernels $\{J_t\}$ is said to be **uniformly mixing** if $\epsilon = \sup_t \epsilon_t > 0$.

Definition 3: [3] The **Birkhoff's contraction coefficient** of any kernel J is, $\tau(J) = \sup_{0 \le h(\mu,\mu') < \infty} \frac{h(J\mu,J\mu')}{h(\mu,\mu')} = tanh[\frac{1}{4}\sup_{\mu,\mu'} h(J\mu,J\mu')]$. h here denotes the Hilbert metric which is defined and explained in [3]. $\tau(J) \le 1$ always and if J is mixing, $\tau(J) \le \tilde{\tau}(J) < 1$ where $\tilde{\tau}(J) \stackrel{\triangle}{=} \frac{1-\epsilon^2}{1+\epsilon^2} < 1$. We denote $\tau(R_t)$ by τ_t and $\epsilon(R_t)$ by ϵ_t. Note that R_t depends on Y_t and hence τ_t and ϵ_t are, in general, random variables.

B. Approximate Non-linear Filtering Using a Particle Filter

The problem of nonlinear filtering is to compute at each time t, the conditional probability distribution, of the state X_t given the observation sequence $Y_{1:t}$, $\pi_t(dx) = Pr(X_t \in dx|Y_{1:t})$. It also evaluates the prediction distribution $\pi_{t|t-1}(dx) = Pr(X_t \in dx|Y_{1:t-1})$. The transition from π_{t-1} to π_t is defined using the Bayes recursion as follows:

$$\pi_{t-1} \text{---}> \pi_{t|t-1} = Q_t(\pi_{t-1}) \text{---}> \pi_t = \frac{\psi_t \pi_{t|t-1}}{(\pi_{t|t-1}, \psi_t)}$$

Now if the system and observation models are linear Gaussian, the posteriors would also be Gaussian and can be evaluated in closed form using a Kalman filter. For nonlinear or nonGaussian system or observation model, except in very special cases, the filter is infinite dimensional. Particle Filtering [10] is a sequential monte carlo technique for approximate nonlinear filtering which was first introduced in [4] as Bayesian Bootstrap Filtering.

A **particle filter** [10] is a recursive algorithm which produces at each time t, a cloud of N particles $\{x_t^{(i)}\}$ whose empirical measure, π_t^N (which is a random measure), closely "follows" π_t. It starts with sampling N times from π_0 to approximate it by $\pi_0^N(dx) \triangleq \frac{1}{N}\sum_{i=1}^N \delta_{x_0^{(i)}}(dx)$. Then for each time step it runs the Bayes recursion which can be summarized as follows:

$$\pi_{t-1}^N \triangleq \frac{1}{N}\sum_{i=1}^N \delta_{x_{t-1}^{(i)}}(dx) \text{---}> \pi_{t|t-1}^N \triangleq \frac{1}{N}\sum_{i=1}^N \delta_{\bar{x}_t^{(i)}}(dx) \text{---}> \bar{\pi}_t^N \triangleq \frac{1}{N}\sum_{i=1}^N w_t^{(i)} \delta_{\bar{x}_t^{(i)}}(dx) \text{---}> \pi_t^N \triangleq \sum_{i=1}^N \delta_{x_t^{(i)}}(dx)$$

$$\text{where} \quad \bar{x}_t^{(i)} \sim Q_t(x_{t-1}^{(i)}, dx), \; x_t^{(i)} \sim \text{Multinomial}(\{\bar{x}_t^{(i)}, w_t^{(i)}\}_{i=1}^N), \quad w_t^{(i)} \triangleq \frac{\psi_t(\bar{x}_t^{(i)})}{(\pi_{t|t-1}^N, \psi_t)}$$

Both $\bar{\pi}_t^N$ and π_t^N approximate π_t but the last step is aimed at reducing the degeneracy of the particles.

III. Change Detection Statistics

A. The ELL statistic

"Expected (negative) Log Likelihood" or ELL [19] at time t, is the *conditional expectation of the negative logarithm of the prior likelihood of the state at time t, under the no change hypothesis (H_0), given observations till time t*, i.e.

$$ELL(Y_{1:t}) \triangleq E[-\log p_t^0(x)|Y_{1:t}] = E_{\pi_t}[-\log p_t^0(x)]. \tag{1}$$

The second equality follows from the definition of π_t, $\pi_t(dx) = Pr(X_t \in dx|Y_{1:t})$. For systems where exact filters do not exist and a PF is used to estimate π_t, the estimate of ELL using the empirical distribution π_t^N becomes $ELL^N = \frac{1}{N}\sum_{i=1}^N[-\log p_t^0(x_t^{(i)})]$. It is interesting to note that ELL as defined above is equal to the Kerridge Inaccuracy [5] between the posterior and prior state pdf.

Definition 4: The **Kerridge Inaccuracy** [5] between two pdfs p and q is defined as $K(p : q) = \int p(x)[-\log q(x)]dx$. It is used in statistics as a measure of inaccuracy between distributions.
We have $ELL(Y_{1:t}) \stackrel{\triangle}{=} E_{\pi_t}[-\log p_t^0(x)] = K(\pi_t : p_t^0)$[4]. Henceforth, we denote $ELL(Y_{1:t}^0) = K(\pi_t^0 : p_t^0) \stackrel{\triangle}{=} K_t^0$ and $ELL(Y_{1:t}^c) = K(\pi_t^c : p_t^0) \stackrel{\triangle}{=} K_t^c$.

Motivation for ELL: The use of ELL (or equivalently Kerridge Inaccuracy) for partially observed systems is motivated by the use of log likelihood for hypothesis testing in the fully observed case. For a fully observed system, one can evaluate $X_t = h_t^{-1}(Y_t)$ from the observation Y_t and then $\log p_t^0(X_t) = \log p_t^0(h_t^{-1}(Y_t))$ would be the log likelihood of the state taking value $X_t = h_t^{-1}(Y_t)$ under H_0. This is proportional to likelihood of Y_t under H_0. If $Y_t = Y_t^0$, then its likelihood (and hence also the likelihood of the state X_t) under H_0 will be larger than if $Y_t = Y_t^c$. But for partially observed systems, X_t is not a deterministic function of $Y_{1:t}$. It is a random variable with distribution π_t. Hence we replace the log likelihood of the state by its expectation under π_t which is the ELL. Note that ELL can also be interpreted as the MMSE of log likelihood of state obtained from the noisy observations.

B. When does ELL work: A Kerridge Inaccuracy perspective

Taking expectation of $ELL(Y_{1:t}^0) = K(\pi_t^0 : p_t^0)$ over normal observation sequences, we get

$$E_{Y_{1:t}^0}[ELL(Y_{1:t}^0)] = E_{Y_{1:t}^0} E_{\pi_t^0}[-\log p_t^0(x)] = E_{p_t^0}[-\log p_t^0(x)] = H(p_t^0) = K(p_t^0 : p_t^0) \stackrel{\triangle}{=} EK_t^0$$

where $H(.)$ denotes entropy. Similarly, for the changed system observations, $E_{Y_{1:t}^c}[ELL(Y_{1:t}^c)] = K(p_t^c : p_t^0) \stackrel{\triangle}{=} EK_t^c$, i.e. the expectation of $ELL(Y_{1:t}^c)$ taken over changed system observation sequences is actually the Kerridge Inaccuracy between the changed system prior, p_t^c, and the original system prior, p_t^0, which will be larger than the Kerridge Inaccuracy between p_t^0 and p_t^0 (entropy of p_t^0) [6].

Now, ELL will detect the change when EK_t^c is "significantly" larger than EK_t^0. Setting the change threshold to

$$\kappa_t \stackrel{\triangle}{=} EK_t^0 + 3\sqrt{VK_t^0}, \text{ where } VK_t^0 = Var_{Y_{1:t}}(K_t^0), \tag{2}$$

will ensure a false alarm probability less than 0.11 (0.05 if unimodal)[5]. By the same logic, if $EK_t^c - 3\sqrt{VK_t^c} > \kappa_t$ then the miss probability [20] (probability of missing the change) will also be less than 0.11 (0.05 if unimodal).

[4]it is actually $K(\frac{d\pi_t}{dx} : p_t^0)$ but as mentioned earlier, we denote the density $\frac{d\pi_t}{dx}$ by the same symbol as the distribution

[5]0.11 follows from the Chebyshev inequality [20]. But if the pdf of $K_t^0(Y_{1:t})$ is unimodal, Gauss's inequality [20] can be applied to show that the probability is less than 0.05

Now evaluating VK_t^0 or VK_t^c analytically is not possible without having an analytical expression for π_t^0 or π_t^c. But we can use Jensen's inequality [21] to bound VK_t^0 (and similarly VK_t^c) as follows: Apply Jensen's inequality on $[-\log p_t(x)]^2$ which is a convex function of $[-\log p_t(x)]$:

$$\begin{aligned}
{K_t^0}^2 &= (E_{\pi_t}[-\log p_t(x)])^2 \leq E_{\pi_t}[[-\log p_t(x)]^2] \\
\text{So, } VK_t^0 = Var_{Y^0_{1:t}}(K_t^0) &= E_{Y_{1:t}}[{K_t^0}^2] - (EK_t^0)^2 \\
&\leq E_{Y_{1:t}}[E_{\pi_t}[[-\log p_t(x)]^2]] - (EK_t^0)^2 = E_{p_t^0}[[-\log p_t^0(x)]^2] - (EK_t^0)^2
\end{aligned}$$

Definition 5: We define a change to be **"detectable"** by ELL (with false alarm and miss probabilities less than 0.11) if $EK_t^c - 3\sqrt{VK_t^c} > \kappa_t$, where $\kappa_t \triangleq EK_t^0 + 3\sqrt{VK_t^0}$.

C. When ELL fails: The OL Statistic

The above analysis assumed no estimation errors in evaluating ELL. But, the PF is optimal for the unchanged system. Hence when estimating π_t (required for evaluating the ELL) for the changed system, there is "exact filtering error". Also the particle filtering error is much larger in this case. The upper bound on the approximation error in estimating the ELL is proportional to the "rate of change" (discussed in Section V). Hence ELL is approximated accurately for a slow change and thus ELL detects such a change as soon as it becomes "detectable" (see definition 5 above in Section III-B). But ELL fails to detect drastic changes because of large estimation error in evaluating π_t. But large estimation error in evaluating π_t also corresponds to a large value of OL (Observation Likelihood) which can be used for detecting such changes (discussed in Theorem 4 in Section VI). OL is the negative log likelihood of the current observation conditioned on past observations under the no change hypothesis, i.e. $OL = -\log Pr(Y_t|Y_{1:t-1}, H_0)$. It is evaluated using a PF as $OL_t^N = -\log(Q_t^0 \pi_{t-1}^N, \psi_t)$. A change is declared if OL exceeds a threshold. Thus for changed observations, $OL_t^{c,0,N} = -\log(Q_t^0 \pi_{t-1}^{c,0,N}, \psi_t^c)$ (notation defined in Section II-A).

OL takes longer to detect a slow change (or may not detect it at all) because of the following reason: Assuming that $\pi_{t-1}^{c,0,N}$ "correctly" approximates π_{t-1}^c (which is true for a slow change), OL uses only the change magnitude at the current time step, $D_{Q,t}$ (defined in Definition 6 of Section V), to detect the change. For a slow change, $D_{Q,t}$ is also small. OL starts detecting the slow change only when the approximation error in $\pi_{t-1}^{c,0,N}$ becomes large enough. This intuitive idea becomes clearer in Theorem 3 in Section V.

D. Defining $p_t(x)$

ELL is given by $E_{\pi_t}[-\log p_t(X)]$ for which we need to know the state prior $p_t(x)$ at each t. Note that we denote $p_t^0(x)$ by $p_t(x)$ in the rest of this paper.

1) For some cases, for e.g. if the state dynamics (or the part of the state dynamics used for detecting change) is linear with Gaussian system noise and Gaussian initial state distribution, $p_t(x)$ can be easily defined in closed form.
2) If $p_t(x)$ of the part of the state vector used to detect the change cannot be defined in closed form for each t, then one solution is to use prior knowledge to define $p_t(x)$ as coming from a certain parametric family for example a Gaussian or a mixture of Gaussians. Its parameters can be learnt using observation noise-free training data sequences. Also if $p_t(x)$ is assumed to be piecewise constant in time, one can use a single training sequence to learn its parameters.

E. Time Averaging

Now single time instant estimates of ELL or OL may be noisy. Hence in practice, we average the statistic over a set of past time frames. Averaging OL over past p frames gives $aOL(p) = \frac{1}{p}[-\log Pr(Y_{t-p+1:t}|Y_{1:t-p})]$. Averaging ELL over past frames is given by $aELL(p) = \frac{1}{p}\sum_{k=t-p+1}^{t} ELL(Y_{1:k})$ but this cannot be justified unless we can show that $ELL(Y_{1:t})$ is ergodic. But one can evaluate joint ELL as $jELL(p,t) = \frac{1}{p}E[-\log p_{t-p+1:t}(X_{t-p+1:t})|Y_{1:t}]$ which is the Kerridge Inaccuracy between the joint posterior distribution of $X_{t-p+1:t}$ given $Y_{1:t}$ and their joint prior. If using $aELL(p,t)$, the threshold $Th(p,t)$ will depend on the sum of individual entropies of $X_{t-p+1:t}$. If using $jELL(p)$, the threshold, $Th(p,t)$, will depend on the joint entropy of $X_{t-p+1:t}$.

Now the value of p [6] can either be set heuristically or one can *modify the CUSUM algorithm [8] to deal with unknown change parameters: Declare a change if*

$$\max_{1\leq p\leq t}[Statistic(p,t) - Th(p,t)] > \lambda. \tag{3}$$

The change time is estimated as $t - p^* + 1$ *where* p^* *is the argument maximizing* $[Statistic(p) - Th(p,t)]$. We have implemented CUSUM on ELL and CUSUM on OL and show results in Section VIII.

[6] Note here that $p = t - j + 1$ (using notation of Section I-C)

IV. ERRORS IN ELL APPROXIMATION

The above analysis for ELL assumes that there are no errors in estimating $ELL(Y^0_{1:t}) = K(\pi^0_t : p_t) \triangleq K^0_t$ and $ELL(Y^c_{1:t}) \triangleq K^c_t$ which is true only if exact finite dimensional filters exist for a problem and correct models for the transition kernel and conditional probability of observation given state are used, e.g. estimation of K^0_t in the linear Gaussian case using a Kalman filter. But in all other cases there are three kinds of errors: When we are trying to estimate K^c_t using the transition kernel for the original system, what we are really evaluating is $K^{c,0}_t \triangleq E_{\pi^{c,0}_t}[-\log p^0_t(x)]$ instead of K^c_t (**"exact filtering error"**). Note that $\pi^{c,0}_t$ is the posterior state distribution for the changed observations estimated using a PF optimal for the unchanged system. We can use stability results from [3] to show that the "exact filtering error" goes to zero (or atleast is monotonically decreasing) for large time instants, for posterior expectations of bounded functions of the state. But $K^{c,0}_t = E_{\pi^{c,0}_t}[-\log p^0_t(x)]$ where $[-\log p^0_t(x)]$ is an unbounded function while the stability results hold only for bounded functions of the state. Considering its bounded approximation introduces **bounding errors** which go to zero as the bound goes to infinity. Also, since we use a PF with a finite number of particles to approximate the optimal filter, there is **PF approximation error**. This error goes to zero as the number of particles goes to infinity.

Now, we quantify our claims. Our aim is to *either* show a result of the type
$\lim_{M\to\infty}(\lim_{N\to\infty} \Xi_{pf}[|K(\pi^0_t : p_t) - K(\pi^{0,N}_t : p^M_t)|]) = 0$ and
$\lim_{M\to\infty}(\lim_{t\to\infty}(\lim_{N\to\infty} \Xi_{pf}[|K(\pi^c_t : p_t) - K(\pi^{c,0,N}_t : p^M_t)|])) = 0, \ a.s.$
where $p^M_t(x) \triangleq \max\{p_t(x), e^{-M}\}$[7] or show that $[-\log p_t(x)]$ is uniformly bounded for all t, so that the outermost convergence with M follows trivially. Under weaker assumptions, we show that even though the error does not converge to zero with time, it is eventually monotonically decreasing with time and hence stable. Note that the analysis of this section can be generalized to the error in evaluating the posterior expectation of any function of the state under the changed system model (not just ELL), when evaluated using a PF that is optimal for the unchanged system model. We use the following two results from [3] to prove our results:

Lemma 1: **("Exact filtering error" bound, Theorem 4.8 of [3])** If for all k, the kernel R_k is a.s. mixing ($\Longrightarrow$ $\epsilon_k > 0, a.s.$ & Birkhoff's contraction coefficient $\tau_k \le \tilde{\tau}_k(\epsilon_k) < 1, a.s.$), then the weak norm between the correct

[7]Note p^M_t is not a pdf.

optimal filter density μ_t and the incorrect one μ'_t is upper bounded as follows:

$$\begin{aligned}
\sup_{\phi:||\phi||_\infty \leq 1} |(\mu_t - \mu'_t, \phi)| &\leq \delta_t + \frac{2\delta_{t-1}}{\epsilon_t^2} + \frac{4}{\log 3} \sum_{k=1}^{t-2} \tilde{\tau}_{t:k+3} \frac{\delta_k}{\epsilon_{k+1}^2 \epsilon_{k+2}^2} &(4)\\
&\triangleq \theta_t(\delta_k, \epsilon_k, 0 \leq k \leq n), a.s. &(5)\\
\text{where} \quad \delta_k &\triangleq \sup_{\phi:||\phi||_\infty \leq 1} |(\mu'_k - \bar{R}_k \mu'_{k-1}, \phi)| \leq 2 &(6)
\end{aligned}$$

Lemma 2: **(PF error bound)**

1) **(Theorem 5.7 of [3])** If for all k, the kernel R_k is a.s. mixing ($\epsilon_k > 0, a.s.$ & $\tau_k \leq \tilde{\tau}_k(\epsilon_k) < 1, a.s.$), and $\sup_{x \in E_{x,y}} \psi_k(x) < \infty, a.s.$, then the weak norm between the correct optimal filter density μ_t and the approximation μ_t^N (evaluated using the PF) is upper bounded as follows:

$$\begin{aligned}
\sup_{\phi:||\phi||_\infty \leq 1} \Xi_{pf}[|(\mu_t - \mu_t^N, \phi)|] &\leq \frac{2(\rho_t + \frac{2\rho_{t-1}}{\epsilon_t^2} + \frac{4}{\log 3} \sum_{k=1}^{t-2} \tilde{\tau}_{t:k+3} \frac{\rho_k}{\epsilon_{k+1}^2 \epsilon_{k+2}^2})}{\sqrt{N}} &(7)\\
&\triangleq \frac{\beta_t(\rho_k, \epsilon_k, 0 \leq k \leq n)}{\sqrt{N}}, a.s. &(8)\\
\text{where} \quad \rho_k &\triangleq \frac{\sup_{x \in E} \psi_k(x)}{\inf_{\mu \in \mathcal{P}(E)} (Q_k \mu, \psi_k)} < \infty, a.s. &(9)
\end{aligned}$$

2) **(Corollary 5.11 of [3])** If the sequence of kernels R_t is uniformly $a.s.$ mixing with t i.e. $\epsilon_k > \epsilon > 0$, then convergence averaged over observations sequences holds uniformly in t, i.e. there exists a $\beta^* < \infty$ s.t. $\sup_{\phi:||\phi||_\infty \leq 1} E_{Y_{1:t}}[\Xi_{pf}[|(\mu_t - \mu_t^N, \phi)|]] < \frac{\beta^*}{\sqrt{N}}$ $\forall t$.

Now we can claim the following results under progressively weaker assumptions. The proofs are given in the Appendix.

Theorem 1: **Asymptotic Stability Results**

1) Assuming (i) Change occurs for only a finite time period $[t_c : t_f]$ and starting time $t_c \leq T^* < \infty$; (ii) $\sup_{x \in E_{x,y}} \psi_k(x) < \infty, a.s.$, $\forall k$; (iii) R_k^c, R_k^0 and $R_k^{c,0} \triangleq Q_k^0(x, dx')\psi_k^c(x')$ are $a.s.$ uniformly mixing with time (i.e. there exists an $\epsilon > 0$ s.t. the mixing parameter $\epsilon_t > \epsilon$ $\forall t$, $a.s.$) and (iv) The posterior state space, $E_{x,Y_t} \triangleq \{x \in E_t : \psi_{t,Y_t}(x) > 0\}$, is a uniformly compact and proper subset of $E_t \triangleq \{x : p_t(x) > 0\}$, then the following result holds:

$$\begin{aligned}
\lim_{N \to \infty} E_{Y_{1:t}}[\Xi_{pf}[|K(\pi_t^0 : p_t) - K(\pi_t^{0,N} : p_t)|]] &= 0, a.s., \text{ uniformly in t}\\
\lim_{t,N \to \infty} E_{Y_{1:t}}[\Xi_{pf}[|K(\pi_t^c : p_t) - K(\pi_t^{c,0,N} : p_t)|]] &= 0, a.s. &(10)
\end{aligned}$$

i.e. $error^c(t, N) \triangleq |K(\pi_t^c : p_t) - K(\pi_t^{c,0,N} : p_t)|$ averaged over PF realizations and observation sequences is asymptotically stable with t for large N [8].

2) Assuming (i), (ii), (iii) as above, a weaker assumption (iv)$'$: Convergence of the error $E_{Y_{1:t}}[|K(\pi_t^c : p_t^M) - K(\pi_t^c : p_t)|]$ to zero as $M \to \infty$ is uniform in t, then we have

$$\begin{aligned} \lim_{M\to\infty}(\lim_{N\to\infty} E_{Y_{1:t}}[\Xi_{pf}[|K(\pi_t^0 : p_t) - K(\pi_t^{0,N} : p_t^M)|]]) &= 0, a.s., \text{ uniformly in t} \\ \lim_{M\to\infty}(\lim_{t,N\to\infty} E_{Y_{1:t}}[\Xi_{pf}[|K(\pi_t^c : p_t) - K(\pi_t^{c,0,N} : p_t^M)|]]) &= 0 \end{aligned} \quad (11)$$

This implies that the $error^c(t, N, M) \triangleq |K(\pi_t^c : p_t) - K(\pi_t^{c,0,N} : p_t^M)|$ averaged over PF realizations and observation sequences is asymptotically stable with t for large N, M.

3) Assuming (i), (ii), (iii) and a weaker assumption (iv)$''$: The posterior state space, $E_{x,Y_t} \triangleq \{x \in E_t : \psi_{t,Y_t}(x) > 0\}$, is a compact and proper subset of $E_t \triangleq \{x : p_t(x) > 0\}$, and (v) increase of $M_t \triangleq \max_{x\in E_{x,Y_t}}[-\log p_t(x)]$ with t is atmost polynomial, then [9] we have

$$\lim_{t\to\infty}(\lim_{N\to\infty} \Xi_{pf}[|K(\pi_t^c : p_t) - K(\pi_t^{c,0,N} : p_t)|]) = 0, a.s. \quad (12)$$

i.e. $[\lim_{N\to\infty}(error^c(t, N))]$ averaged over PF realizations is a.s. asymptotically stable with t [10].

Proof: See Appendix ■

The assumption (iv) in Theorem 1.1 implies that $[-\log p_t(x)]$ is uniformly bounded $\forall x$ in the support set of π_t, π_t^c, $\forall t$, so that Lemmas 1 and 2 can be directly applied to prove the result. But one can relax this assumption (in Theorem 1.2) by defining a sequence of increasing functions $\{[-\log p_t^M(x)]\}$ with $p_t^M(x) = \max\{p_t(x), e^{-M}\}$, s.t. $\lim_{M\to\infty}[-\log p_t^M(x)] = [-\log p_t(x)]$. Then by a simple extension of Monotone Convergence Theorem ([7], page 87) to functions which could be negative but are bounded from below, we have $\lim_{M\to\infty} K(\pi_t^c : p_t^M) = K(\pi_t^c : p_t)$. We then get Theorem 1.2 which requires the assumption that this convergence be uniform in t. It is difficult to show the convergence with M holding uniformly for all t, almost surely over all observation sequences (since π_t is not available in closed form). But it is easy to show for a large class of general HMMs that the assumption is satisfied in mean over observation sequences (see the example of Section VII-A). Using this assumption, the convergence result in Theorem 1.2 is also a 'convergence in the mean' result.

[8] This means the following: For every $\epsilon > 0$, there exists an N^* and a T^* (N^* does not depend on T^*) s.t. $\forall N > N^*$ and $\forall t > T^*$, $E_{Y_{1:t}}[\Xi_{pf}[error^c(t, N)]] < \epsilon$. Also note that for normal observations, the exact filtering error is itself zero (hence asymptotic stability with t is meaningless)

[9] Result for normal observations is same as in (11)

[10] This means the following: For every $\epsilon > 0$, there exists a T^* s.t. $\forall t > T^*$, $\lim_{N\to\infty}(\Xi_{pf}[error(t, N)] < \epsilon$ (or that for every $t > T^*$, there exists an N^* which depends on t and ϵ, s.t. for all $N > N^*$, $error(t, N) < 2\epsilon$)

One can also relax the assumption (iv) of Theorem 1.1 in a different way, as in Theorem 1.3. Here we assume that the posterior state space is compact for each t and assume that the increase of M_t (the bound on $[-\log p_t(x)]$) is atmost polynomial. Under this assumption, one can show asymptotic stability of the errors, but in this case a different N is required for each t (convergence with N is not uniform in t).

If the unnormalized filter kernels, R_k^c, R_k^0 and $R_k^{c,0}$, are mixing (but not uniformly mixing), convergence of the error to zero (asymptotic stability with time) will not hold. But we can still claim eventual monotonic decrease (and hence stability) of the error with time. We have the following results for changed observations. Note that even under this weaker assumption, the results for normal observations remain the same as in Theorem 1, except that the convergence with N is not uniform with t:

Theorem 2: **Stability Results**

1) Assuming (i), (ii), a weaker assumption (iii)$'$: R_k^c, R_k^0 and $R_k^{c,0}$ are mixing and (iv)$'$: Convergence of the error $E_{Y_{1:t}}[|K(\pi_t^c : p_t^M) - K(\pi_t^c : p_t)|]$ to zero as $M \to \infty$ is uniform in t (as in Theorem 1.2), we have the following result: Given any $\Delta > 0$, there exists an M_Δ s.t.

$$\lim_{N\to\infty} E_{Y_{1:t}}[\Xi_{pf}[|K(\pi_t^c : p_t) - K(\pi_t^{c,0,N} : p_t^{M_\Delta})|]] \quad \leq \quad \Delta + M_\Delta E_{Y_{1:t}}[\theta_t^{c,0}] \tag{13}$$

where $\theta_t^{c,0} \stackrel{\triangle}{=} \theta_t(\delta_k^{c,0}, \epsilon_k^c, t_c \leq k \leq t)$. $\theta_t^{c,0}$ and hence also $E_{Y_{1:t}}[\theta_t^{c,0}]$ is eventually monotonically decreasing with time. It is easy to see that this implies that $\lim_{N\to\infty} E_{Y_{1:t}}[\Xi_{pf}[error]]$ is eventually monotonically decreasing with t and hence stable.

2) Assuming (i), (ii), (iii)$'$ and (iv)$''$: The posterior state space, $E_{x,Y_t} \stackrel{\triangle}{=} \{x \in E_t : \psi_{t,Y_t}(x) > 0\}$, is a compact and proper subset of $E_t \stackrel{\triangle}{=} \{x : p_t(x) > 0\}$ (as in Theorem 1.3), we have

$$\lim_{N\to\infty} \Xi_{pf}[|K(\pi_t^c : p_t) - K(\pi_t^{c,0,N} : p_t)|] \quad \leq \quad M_t \theta_t^{c,0} \tag{14}$$

where $\theta_t^{c,0}$ eventually monotonically decreases with time. It is easy to see that this implies that $\frac{\lim_{N\to\infty} \Xi_{pf}[error]}{M_t}$ is eventually monotonically decreasing with t and hence stable.

Proof: See Appendix ■

V. Effect of Increasing Rate of Change on Approximation Errors

Since the aim is to detect a change as soon as possible and with a given finite number of particles, we need to study the finite time, finite number of particles behavior of the bounds obtained in the previous section. ELL will detect the change, if its approximation exceeds the detection threshold (inspite of the errors). Applying theorem 2.2, we have

$\Xi_{pf}[|K_t^0 - K_t^{0,M,N}|] < \frac{M_t \beta_t^0}{\sqrt{N}}$ and
$\Xi_{pf}[|K_t^c - K_t^{c,0,M,N}|] < \frac{M_t \beta_t^{c,0}}{\sqrt{N}} + M_t \theta_t^{c,0} \triangleq e_t^{c,0,N}$
where $\beta_t^0 = \beta_t(\rho_k^0, \epsilon_k^0, 0 \le k \le t)$, $\theta_t^{c,0} = \theta_t(\delta_k^{c,0}, \epsilon_k^c, t_c \le k \le t)$, and $\beta_t^{c,0} = \beta_t(\rho_k^0, \epsilon_k^0, 0 \le k \le t_c, \rho_k^{c,0}, \epsilon_k^{c,0}, t_c \le k \le t)$ and θ_t, β_t defined in (5), (8) respectively. Thus for ELL to detect a change, we need to show that $K_t^c - M_t \theta_t^{c,0} - \frac{M_t \beta_t^{c,0}}{\sqrt{N}}$ exceeds the detection threshold.

We show in this section that the "exact filtering error" bound, $\theta_t^{c,0}$, and the PF error bound coefficient, $\beta_t^{c,0}$ (and hence also the total error, $e_t^{c,0,N}$) are upper bounded by increasing functions of the "rate of change" metric (defined below) with increasing derivatives of all orders. We also show that the observation likelihood, OL, is upper bounded by an increasing function of the "rate of change" metric. Note that although we prove the above result for the error in ELL estimation, it can directly generalize to bounding the error between the true posterior expectation of any function of the state and its posterior expectation estimated by a PF with incorrect system model assumptions. The "rate of change" metric can be generalized to a metric for system model error per time step.

We give below some definitions and then state a sequence of lemmas required to prove the main result.

Definition 6: We define a **distance metric between state transition kernels** Q_t^c **and** Q_t^0 (a metric for the **rate of change**), for a given observation Y_t, $D_{Q,Y_t}(Q_t^c, Q_t^0)$, as the following distance between R_{t,Y_t}^c, R_{t,Y_t}^0:

$$\begin{aligned} D_{Q,Y_t}(Q_t^c, Q_t^0) &\triangleq D_R(R_{t,Y_t}^c, R_{t,Y_t}^0) \\ &\triangleq \sup_x \int_E |R_{t,Y_t}^c(x,x') - R_{t,Y_t}^0(x,x')| dx' \\ &= \sup_x \int_E \psi_{t,Y_t}(x') |Q_t^c(x,x') - Q_t^0(x,x')| dx' \end{aligned}$$

It is easy to show that, for a given observation Y_t, D_R and hence D_Q satisfy the properties of a metric over the space of transition kernels. We use $D_{Q,t}$ to denote $D_{Q,Y_t^c}(Q_t^c, Q_t^0)$ for ease of notation.

Definition 7: We define the **vector of rates of change**, $\underline{D_Q}$ as

$$\underline{D_Q} \triangleq [D_{Q,t_c}, ...D_{Q,k}, ...D_{Q,t_f}] \tag{15}$$

Definition 8: The **total exact filtering error in the posterior** is defined as the total variation norm of the difference between the posteriors evaluated using the correct and the incorrect model, scaled by $\lambda_{k,Y_k^c}^c(E)$ where $\lambda_{k,Y_k^c}^c$ is the invariant measure [3] corresponding to $R_{k,Y_k^c}^c$ [11],

$$\tilde{D}_{t,Y_{0:t}} \triangleq \lambda_{k,Y_k^c}^c(E) ||\pi_t^{c,0} - \pi_t^{c,c}||. \tag{16}$$

$\tilde{D}_{t,Y_{0:t}}$ is a temporary variable used to write the lemmas more clearly.

[11] We scale by $\lambda_{k,Y_k^c}^c(E)$ only for ease of notation in stating theorems

Definition 9: We say that a function $\alpha(z)$ belongs to the **"Alpha functions' class"** if it is an increasing function of z and its derivatives w.r.t. z of all orders are also increasing functions. Note z can be a scalar or a vector but $\alpha(z)$ is a scalar.

We state here a lemma for the Alpha functions' class[12] which we use to prove later lemmas.

Lemma 3: **(Composition Lemma):** The composition of two Alpha functions is also an Alpha function, i.e. if $\alpha_1(x,z), \alpha_2(y)$ are Alpha functions of their arguments, then their composition function $\alpha_1(x, \alpha_2(y))$ is also an Alpha function of $[x, y]$.

Proof: See Appendix ■

Now, we need to show that θ_t, β_t are upper bounded by Alpha functions of $\underline{D_Q}$. This will follow if we can show a similar result for $\delta_k,\ \rho_k,\ \forall k \geq t_c$. We first show in Lemma 4 that $\delta_k,\ \rho_k$ are upper bounded by Alpha functions of $[D_{Q,k}, \tilde{D}_{k-1}]$ (first proved in [22]). Then in Lemma 5, we use a mathematical induction argument and Lemma 3 to show that $\tilde{D}_{k-1}$ and δ_k are upper bounded by an Alpha function of $\underline{D_Q}$ for all k. The Alpha function bound on ρ_k (in Lemma 5) follows from the Alpha function bound on $\tilde{D}_{k-1}$ and Lemma 3.

Lemma 4: Defining

$$A_k \triangleq R^c_{k,Y^c_k}(\pi^{c,0}_{k-1})(E), \quad \text{and} \quad C \ \triangleq \ R^c_{k,Y^c_k}(\pi^c_{k-1})(E) \tag{17}$$

$$\text{and assuming} \quad C > \frac{(\tilde{D}_{k-1})}{\epsilon^c_k} + D_{Q,k},\ \forall k, \tag{18}$$

δ_k and ρ_k are upper bounded by Alpha functions of $[D_{Q,k}, \tilde{D}_{k-1}, \frac{1}{\epsilon^{c,0}_k}]$[13], i.e.

$$\delta_k \ \leq \ \frac{2D_{Q,k}}{A_k} \leq \frac{2D_{Q,k}}{C - \frac{\tilde{D}_{k-1}}{\epsilon^c_k}} \triangleq \tilde{\alpha}_{\delta,k}([D_{Q,k}, \tilde{D}_{k-1}]) \tag{19}$$

$$\rho_k \ \leq \ \frac{\sup_x \psi_{k,Y_k}(x)}{{\epsilon^{c,0}_k}^2 (A_k - D_{Q,k})} \leq \frac{\sup_x \psi_{k,Y_k}(x)}{{\epsilon^{c,0}_k}^2 (C - \frac{\tilde{D}_{k-1}}{\epsilon^c_k} - D_{Q,k})} \triangleq \tilde{\alpha}_{\rho,k}([D_{Q,k}, \tilde{D}_{k-1}, \frac{1}{\epsilon^{c,0}_k}]),\ a.s. \tag{20}$$

Proof: See Appendix ■

Lemma 5: Assuming (18) holds, $\tilde{D}_t$, δ_t and ρ_t are upper bounded by Alpha functions of $\underline{D_Q}$, i.e.

$$\tilde{D}_t \ \leq \ \alpha_{\tilde{D},t}(\underline{D_Q}),\ \forall t \geq t_c \tag{21}$$

$$\delta_t \ \leq \ \alpha_{\delta,t}(\underline{D_Q}),\ \forall t \geq t_c \tag{22}$$

$$\rho_t \ \leq \ \alpha_{\rho,t}(\underline{D_Q}, \frac{1}{\epsilon^{c,0}_t}),\ \forall t \geq t_c \tag{23}$$

[12]We are not sure if this class of functions the composition lemma given below already exist in literature

[13]Note that ϵ^c_k is not a function of the rate of change and hence we treat it as a constant in this entire analysis

Proof: We use mathematical induction to prove (21) and (22). (23) then follows from (20) of Lemma 4, (21) and Lemma 3. First note that $\tilde{D}_t = 0 = \delta_t, \ \forall t < t_c$. The base case, $t = t_c$, is true since

$$\delta_{t_c} \quad \leq \quad \frac{2D_{Q,t_c}}{C} \triangleq \alpha_{\delta,t_c}(\underline{D_Q}) \tag{24}$$

$$\tilde{D}_{t_c} \quad = \quad ||\pi_{t_c}^{c,0} - \bar{R}_{t_c}^c \pi_{t_c-1}^0|| = ||\pi_{t_c}^{c,0} - \bar{R}_{t_c}^c \pi_{t_c-1}^{c,0}|| \leq \frac{2D_{Q,t_c}}{C} \triangleq \alpha_{\tilde{D},t_c}(\underline{D_Q}) \tag{25}$$

Inequality (24) follows from (19) of Lemma 4 by putting $\tilde{D}_{t_c-1} = 0$. The last inequality in (25) follows by applying (59) from Appendix with $\tilde{D}_{t_c-1} = 0$.

Now, assume that (21) and (22) hold for $t_c \leq k \leq (t-1)$, i.e. assume that

$$\tilde{D}_{t-1} \quad \leq \quad \alpha_{\tilde{D},t-1}(\underline{D_Q}) \tag{26}$$

$$\text{and} \quad \delta_k \quad \leq \quad \alpha_{\delta,k}(\underline{D_Q}), \ \forall t_c \leq k \leq (t-1). \tag{27}$$

By (19) of Lemma 4, this implies that

$$\delta_t \leq \frac{2D_{Q,t}}{C - \frac{\alpha_{\tilde{D},t-1}(\underline{D_Q})}{\epsilon_t}} \triangleq f(\underline{D_Q}) \tag{28}$$

Now it is easy to see that $f(\underline{D_Q}) = \alpha_1(\underline{D_Q}, \alpha_2(\underline{D_Q}))$ is a composition of two Alpha functions, $\alpha_1(\underline{D_Q}, z) = \frac{2D_{Q,t}}{C-z}$ [14] and $\alpha_2(\underline{D_Q}) = \alpha_{\tilde{D},t-1}(\underline{D_Q})$. Using Lemma 3 (Composition Lemma), the composition of two Alpha functions is also an Alpha function. Thus, $f(\underline{D_Q}) = \alpha_{\delta,t}(\underline{D_Q})$. Now, by Theorem 4.6 of [3], we have that

$$\tilde{D}_t \leq \delta_t + \frac{\delta_{t-1}}{\epsilon_t^{c2}} + \sum_{k=t_c}^{t-2} \tilde{\tau}_{t:k+2} \frac{\delta_k}{{\epsilon_{k+1}^c}^2} \tag{29}$$

Also, we have from (27) and (28) that each of the $\delta_k, k = t_c, ...t$, is upper bounded by an Alpha function. Hence it is easy to see that $\tilde{D}_t$ is also upper bounded by an Alpha function, $\alpha_{\tilde{D},t} \triangleq \alpha_{\delta,t} + \frac{\alpha_{\delta,t-1}}{\epsilon_t^{c2}} + \sum_{k=t_c}^{t-2} \tilde{\tau}_{t:k+2} \frac{\alpha_{\delta,k}}{{\epsilon_{k+1}^c}^2}$. Thus we have proved that (21) and (22) hold for t given that they hold for all $t_c < k \leq (t-1)$. We showed the base case, $t = t_c$, in (24) and (25). Hence by Mathematical Induction, (21) and (22) hold for all $t \geq t_c$. The third equation, (23), follows directly by combining (20), (21) and Lemma 3. ■

The main result of this section given below follows as a corollary of the above lemmas.

Theorem 3: **("Rate of Change" bound)** Assuming the inequality (18), the following results hold:

1) Both the "exact filtering error", $\theta_t(\delta_k, \epsilon_k^c, t_c \leq k \leq t)$, and the PF approximation error coefficient, $\beta_t(\rho_k, \epsilon_k^{c,0}, 0 \leq k \leq t)$, are upper bounded by Alpha functions of the vector of rates of change, $\underline{D_Q}$, and

[14] It is easy to see that $\frac{2}{(C-z)}$ is an Alpha function

consequently the total error $e_t^{c,0,N} = M_t\theta_t^{c,0} + \frac{M_t\beta_t^{c,0}}{\sqrt{N}}$ is also upper bounded by an Alpha function of $\underline{D_Q}$. Also $e_t^{c,0,N}$ increases with t as long as the change persists.

2) The observation likelihood is upper bounded by an **increasing function** (note, it is not an Alpha function) of the vector of rates of change, $\underline{D_Q}$, i.e.

$$OL_t^{c,0} \leq -\log(A_t - D_{Q,t}) \leq -\log(C - \frac{\tilde{D}_{t-1}}{\epsilon_t^c} - D_{Q,t}) \leq -\log(C - \alpha_{\tilde{D},t-1}(\underline{D_Q}) - D_{Q,t}) \tag{30}$$

Proof: Theorem 3.1 follows from the definitions of θ_t, β_t (equations (5) and (8)), Lemma 5 and the following two facts: (a) ϵ_k^c is independent of $D_{Q,k}$ and (b) $\epsilon_k^{c,0}$ is a decreasing function of the rate of change (We do not have a proof for this in the general case). The intuition is that with increasing rate of change, the overlap between Y_k^c and the spread of Q_k^0 decreases and so the kernel $R_k^{c,0}$ becomes less mixing ($\epsilon_k^{c,0}$ decreases). In Theorem 3.2, the first inequality of (30) follows by applying (57) (in Appendix), the second one follows by (58) (in Appendix) and the third inequality follows from (21). ■

Thus we have shown that a small rate of change implies that $OL^{c,0}$ is small and hence OL does not detect the change. But it also implies that ELL estimation error, $e_t^{c,0,N}$, is small, which implies that ELL will detect the change as soon as it becomes "detectable" (defined in Definition 5).

The Alpha function nature of the bound on ELL approximation error implies that ELL *is approximated accurately for slow changes, and for some time (until total change magnitude is small) but the error bound blows up quickly to infinity with increasing rate of change ($D_{Q,k}$) or increasing total change magnitude $\tilde{D}_{k-1}$*. Two possible implications of this are[15]: (a) A sequence of small changes introduces less total error (less total error upper bound) than one drastic change of the same total magnitude. (b) Using a PF state transition kernel Q_k^{pf} that may not be equal to the unchanged state transition kernel Q_k^0 but if its distance from Q_k^c is smaller than the distance of Q_k^0 from Q_k^c may introduce less total error (error upper bound). For e.g., setting the PF system noise variance to a larger value than that of Q_k^0 reduces $D_Q(Q_t^{pf}, Q_t^c)$ even though $D_Q(Q_t^{pf}, Q_t^c) \neq 0$ and doing this reduces the total approximation error. This idea has been used in past works [4], [10].

VI. Complementary behavior of ELL and OL

We quantify the complementary behavior of ELL and OL by bounding the ELL approximation error by an increasing function of OL. First consider the PF error coefficient, $\beta_t^{c,0}$. It depends on past values of $\rho_k^{c,0}$ and on

[15] We are comparing error upper bounds in inferring these two facts and hence these inferences may not necessarily hold.

$\epsilon_k^{c,0}$. Using Remark 5.10 of [3], we have the following upper and lower bounds on ρ_k which can be expressed in terms of $OL_k^{c,0}$:

$$\begin{aligned} \frac{\sup_{x\in E_{x,Y_t}}\psi_k^c(x)}{(Q_k^0\pi_{k-1}^{c,0},\psi_k^c)} &\leq \rho_k^{c,0} \leq \frac{\sup_{x\in E_{x,Y_t}}\psi_k^c(x)}{(\epsilon_k^{c,0})^2(Q_k^0\pi_{k-1}^{c,0},\psi_k^c)} \\ \implies \frac{\sup_{x\in E_{x,Y_t}}\psi_k^c(x)}{e^{-OL_k^{c,0}}} &\leq \rho_k^{c,0} \leq \frac{\sup_{x\in E_{x,Y_t}}\psi_k^c(x)}{(\epsilon_k^{c,0})^2 e^{-OL_k^{c,0}}} \end{aligned} \tag{31}$$

Now consider the "exact filtering error" bound, $\theta_t^{c,0}$. It depends on past values of $\delta_k^{c,0}$ and ϵ_k^c. We use inequality (6) of [3] which states that

$$|\bar{\mu}-\bar{\mu}'| \leq \frac{||\mu-\mu'||}{\mu(E)} + \frac{|\mu(E)-\mu'(E)|}{\mu(E)}. \tag{32}$$

Taking $\mu = R_k^{c,0}(\pi_{k-1}^{c,0})$ and $\mu' = R_k^{c,c}(\pi_{k-1}^{c,0})$, using the fact that $R_k^{c,0}(\pi_{k-1}^{c,0})(E) = e^{-OL_k^{c,0}}$ and using inequalities (56) and (57) from the Appendix, we can bound $\delta_k^{c,0}$ in terms of $OL_k^{c,0}$ as:

$$\delta_k^{c,0} \leq \frac{2D_{Q,k}}{e^{-OL_k^{c,0}}} \tag{33}$$

Thus we have the following theorem:

Theorem 4: **(ELL-OL Complementariness)**

1) The ELL approximation error at time t, $e_t^{c,0,N} \triangleq M_t\theta_t^{c,0} + \frac{M_t\beta_t^{c,0}}{\sqrt{N}}$ is upper bounded by an increasing function of past values of $OL_k^{c,0}$ and past values of $D_{Q,k}$, $\frac{1}{\epsilon_k^{c,0}}$, i.e.

$$e_t^{c,0,N} \leq \sum_{k=t_c}^{t} e^{OL_k^{c,0}}\omega_k(\frac{1}{\epsilon_k^{c,0}}, D_{Q,k}) \tag{34}$$

where ω_k is an increasing function of its arguments and is defined by upper bounding $\theta_t^{c,0}$ and $\beta_t^{c,0}$ defined in (5) and (8) respectively using the bounds given in (31) and (33) respectively.

2) The PF error upper bound coefficient is lower bounded by an increasing function of $OL_k^{c,0}$, i.e.

$$\beta_t^{c,0} \geq \sum_{k=t_c}^{t} e^{OL_k^{c,0}}(\sup_{x\in E_{x,Y_k}}\psi_k^c(x))\tilde{\omega}(\frac{1}{\epsilon_k^{c,0}}) \tag{35}$$

Proof: The proof of Theorem 4.1 follows directly by combining the definitions of $\theta_t^{c,0}$ and $\beta_t^{c,0}$ given in (5) and (8) with (31) and (33). Proof of 4.2 follows directly from (31). ■

Now, if a certain change is not detected by OL until time t, it means that all values of OL, $OL_{t_c}^{c,0}, ...OL_k^{c,0}, ...OL_t^{c,0}$ are small (below threshold). This implies, by the above theorem, that the bound on the ELL approximation error is also small or that ELL is approximated accurately. Thus the change will get detected by ELL once its magnitude becomes large enough to satisfy the "detectability" condition (definition 5 in Section III). Conversely, if ELL does not detect a change that is "detectable", it means that the ELL approximation error is large. By the above theorem

(Theorem 4.1) this implies that at least one of $OL_{t_c}^{c,0},...OL_{k}^{c,0},...OL_{t}^{c,0}$ is large and hence OL will detect the change. Thus, we *propose to use a combination of ELL and OL to detect a change when the rate of change can be slow or fast and change parameters are unknown*. A change should be declared when either ELL or OL exceed their respective threshold.

VII. Discussion

A. An Example

Consider the case where Q_t^0, Q_t^c and π_0 are linear Gaussian, so that p_t^0 and p_t^c are also Gaussian. Assume scalar state and observation and let π_0 be zero mean with zero variance. Let the pdf of $Q_t(x, dx')$ is $\mathcal{N}(x, \sigma_{sys}^2)$ and pdf of $Q_t^c(x, dx')$ is $\mathcal{N}(x+\Delta a, {\sigma_{sys}^c}^2)$ with $\sigma_{sys}^c = 0.25\sigma_{sys}$. Also assume that the changed system model lasts for a finite time $[t_c, t_f]$. Thus $p_t^0(x)$ is $\mathcal{N}(0, \sigma_t^2)$ with $\sigma_t^2 = t\sigma_{sys}^2$ and $p_t^c(x)$ is $\mathcal{N}(a_t, \sigma_t^{c2})$ with $a_t = 0,\ \sigma_t^{c2} = t\sigma_{sys}^2,\ \forall t < t_c$, $a_t = (t - t_c + 1)\Delta a,\ \sigma_t^{c2} = t_c\sigma_{sys}^2 + (t - t_c + 1){\sigma_{sys}^c}^2,\ \forall t_c \le t \le t_f$ and $a_t = a_{t_f},\ \sigma_t^{c2} = t_c\sigma_{sys}^2 + (t_f - t_c + 1){\sigma_{sys}^c}^2 + (t - t_f)\sigma_{sys}^2\ \forall t > t_f$. Thus even though the change lasts for a finite time, its effect on $p_t(x)$ is permanent ($p_t^c(x)$ has mean a_{t_f} $\forall t > t_f$). We consider a simple observation model $Y_t = h(X_t) + w_t$ with $h(x) = x^3$. We let w_t be truncated Gaussian observation noise with variance σ_{obs}^2 and truncation parameter, $B < \infty$.

1) Theorem 2.2 holds: A truncated Gaussian observation noise, and the fact that h^{-1} is continuous, makes the support set of $\psi_k(x)$ compact. By the argument given in Example 3.10 of [3] (explained in Section VII-B), this along with the fact that π_0 has finite support (here, zero support) makes the unnormalized filter kernels, $R_t^0, R_t^{c,0}, R_t^c$, mixing, even though the state transition kernels Q_t^0, Q_t^c are not mixing. Also, $\sup_{x \in E_{x,Y_t}} \psi_k(x) = \frac{1}{\sqrt{2\pi}\sigma_{obs}} < \infty$ and the change lasts for a finite time. Also, $M_t = \sup_{x \in E_{x,Y_t}}[-\log p_t(x)] = \sup_{x \in h^{-1}([Y_t - B, Y_t + B])}[-\log p_t(x)] = -\log p_t((|Y_t| + B)^{1/3})$. Thus we satisfy all assumptions for Theorem 2.2.

2) Theorem 2.1 holds: We can show that this example satisfies assumption (iv)$'$ and hence Theorem 2.1 also holds. This is shown as follows: Consider $E_{Y_{1:t}}[|K(\pi_t^c : p_t) - K(\pi_t^c : p_t^M)|]$. By definition of p_t^M, $K(\pi_t^c : p_t) > K(\pi_t^c : p_t^M)$ $\forall Y_{1:t}$ and so

$$\begin{aligned} E_{Y_{1:t}}[|K(\pi_t^c : p_t) - K(\pi_t^c : p_t^M)|] &= E_{Y_{1:t}}[K(\pi_t^c : p_t) - K(\pi_t^c : p_t^M)] = E_{Y_{1:t}}[K(\pi_t^c : p_t)] - E_{Y_{1:t}}[K(\pi_t^c : p_t^M)] \\ &= K(p_t^c : p_t) - K(p_t^c : p_t^M) \triangleq err(M, t) \end{aligned} \tag{36}$$

Here p_t^c and p_t are both Gaussian and hence $err(M, t)$ simplifies to (w.l.o.g. assume $a_t > 0$)

$$err(M, t) = K(p_t^c : p_t) - K(p_t^c : p_t^M) = 2\int_{\sqrt{M}\sigma_t}^{\infty} \frac{x^2}{2\sigma_t^2}\frac{1}{\sqrt{2\pi}\sigma_t^c} e^{-\frac{(x-a_t)}{2\sigma_t^{c2}}} dx \tag{37}$$

Set $y = \frac{x-a_t}{\sigma_t^c}$ and use the fact that $\frac{\sigma_t^c}{\sigma_t} < 1 \ \forall t$.

$$err(M,t) = \int_{\sqrt{M}\frac{\sigma_t}{\sigma_t^c}-\frac{a_t}{\sigma_t^c}}^{\infty} (y\frac{\sigma_t^c}{\sigma_t} + \frac{a_t}{\sigma_t})^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy \quad \leq \quad \int_{\sqrt{M}\frac{\sigma_t}{\sigma_t^c}-\frac{a_t}{\sigma_t^c}}^{\infty} (y^2 + \frac{a_t^2}{\sigma_t^2} + 2\frac{a_t y}{\sigma_t}) \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy$$
$$\leq \quad \int_{\sqrt{M}-\frac{a_t}{\sigma_t^c}}^{\infty} (y^2 + \frac{a_t^2}{{\sigma_t}^2} + 2\frac{a_t y}{\sigma_t}) \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy$$

Now the above is an increasing function of a_t and a decreasing function of σ_t and of σ_t^c. Also, we know that $a_t = a_{t_f} \ \forall t > t_f$ and so $a_t \leq a_{t_f} \ \forall t$. Also, $\sigma_t \geq \sigma_1 = \sigma_{sys} \ \forall t$ and $\sigma_t^c \geq \sigma_1 = \sigma_{sys} \ \forall t$. Thus we have

$$err(M,t) \leq \int_{\sqrt{M}-\frac{a_{t_f}}{\sigma_{sys}}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} (y^2 + \frac{a_{t_f}^2}{{\sigma_{sys}}^2} + 2\frac{a_{t_f} y}{\sigma_{sys}}) dy \triangleq err^*(M) \tag{38}$$

Now $err^*(M)$ is independent of t and $\lim_{M\to\infty} err^*(M) = 0$. Thus assumption (iv)′ is true and hence all assumptions for Theorem 2.1 are satisfied.

3) Performance analysis: "Detectability" condition: Now first assume that no errors are present and do the analysis of Section III-B to find the detection delay. Assume $t_c \approx 0$ to simplify expressions. Then,

$$EK_t^0 = K(p_t^0 : p_t^0) \quad = \quad 0.5 \log 2\pi\sigma_t^2 + 0.5 \tag{39}$$

$$\begin{aligned} EK_t^c = K(p_t^c : p_t^0) \quad &= \quad 0.5 \log 2\pi\sigma_t^2 + 0.5\frac{{\sigma_t^c}^2 + a_t^2}{\sigma_t^2} \\ &\approx \begin{cases} 0.5 \log 2\pi\sigma_t^2 + 0.031 + 0.5\frac{a_t^2}{t\sigma_{sys}^2}, & t \leq t_f \\ 0.5 \log 2\pi\sigma_t^2 + 0.5[0.062\frac{t_f}{t} + \frac{t-t_f}{t}] + 0.5\frac{a_{t_f}^2}{t\sigma_{sys}^2} & t > t_f \end{cases} \end{aligned} \tag{40}$$

$$VK_t^0 \quad \leq \quad E_{p_t^0}[[-\log p_t^0(x)]^2] - (EK_t^0)^2 = 0.5 \tag{41}$$

$$\begin{aligned} VK_t^c \quad &\leq \quad E_{p_t^c}[[-\log p_t^0(x)]^2] - (EK_t^c)^2 = 0.5\frac{{\sigma_t^c}^4}{\sigma_t^4} + \frac{{\sigma_t^c}^2 a_t^2}{\sigma_t^4} \\ &\approx \begin{cases} 0.002 + \frac{0.062 a_t^2}{t\sigma_{sys}^2} & t \leq t_f \\ 0.5[0.062\frac{t_f}{t} + \frac{t-t_f}{t}]^2 + [0.062\frac{t_f}{t} + \frac{t-t_f}{t}]\frac{a_{t_f}^2}{t\sigma_{sys}^2} & t > t_f \end{cases} \end{aligned} \tag{42}$$

The threshold $\kappa_t = EK_t^0 + 3\sqrt{VK_t^0} \leq 2.62$. Set $\kappa_t = 2.62$. The mean distance of K_t^c from the threshold (assuming $t_c \approx 0$) is then:

$$\gamma_t \triangleq EK_t^c - \kappa_t = \begin{cases} 0.5t\frac{(\Delta a)^2}{{\sigma_{sys}}^2} - 2.59 & t_c \leq t \leq t_f \\ \frac{0.5(a_{t_f})^2}{t{\sigma_{sys}}^2} - 2.62 + 0.5[\frac{0.062 t_f}{t} + \frac{t-t_f}{t}] & t > t_f \end{cases} \tag{43}$$

Now consider $t \leq t_f$. We can then apply definition 5 to infer the following: Assuming no approximating errors, the miss probability at time t will be less than 0.11 (0.05 if unimodal) if $\gamma_t > 3\sqrt{VK_t^c}$ which simplifies to $0.5r^2 - 2.59 > .75r$ with $r = a_t/\sigma_t$. It is easy to see that this equation is satisfied for $r \geq 3.2$. Since $t_c \approx 0$, $r \approx \frac{\sqrt{t}\Delta a}{\sigma_{sys}}$. This implies that if the rate of change is of the order of system noise, $\Delta a = \sigma_{sys}$, then with probability greater than 0.89, the change will get detected in $(3.2)^2 = 10.24$ time units or more. This of course is obtained using loose bounds (loose variance bound and the loose Chebyshev or Gauss's inequality bound) and in practice changes can get detected much faster if there are no approximation errors. Infact even with approximation errors, which tend to reduce the value of ELL[16], we see in simulations that the change $\Delta a = \sigma_{sys}$, gets detected faster than this (see Figure 1(a)).

4) Performance analysis: Effect of Approximation Errors: Now we analyze the effect of approximation errors. Applying definition 5 while taking into account the approximation errors, we get: A change will get detected w.p. greater than 0.89, if $\gamma_t - M_t\theta_t^{c,0} > 3\sqrt{VK_t^c}$ (assuming $\frac{M_t\beta_t}{\sqrt{N}}$ can be made small enough by taking N large enough). γ_t is defined in (43). Now, $\delta_k = 0, \forall\ k < t_c,\ k > t_f$. For simplicity, assume $\delta_k = \delta, \forall\ t_c \leq k \leq t_f$, then we have for $t_c \leq t \leq t_f$,

$$\theta_t = \delta + \frac{\delta}{(\epsilon_t^c)^2} + \sum_{k=t_c}^{t-2} (\tau_k)^{(t-k-2)} \frac{\delta}{(\epsilon_k^c)^4}. \tag{44}$$

From (43) and (44), we see that both γ_t and θ_t increase till t_f. γ_t has an approximately linear increase (for small t_c), $\Delta\gamma_t \approx \frac{0.5(\Delta a)^2}{{\sigma_{sys}}^2} = 0.5$, while θ_t increases at decreasing rates of increase[17], $\Delta\theta_t = \tau_t \Delta\theta_{t-1}$. Now, if the change is slow enough so that $\frac{M_t\delta}{\epsilon^4} < 0.5$, then $\gamma_t - M_t\theta_t$ will increase with time until t_f. The change will get detected when $\gamma_t - M_t\theta_t$ exceeds zero. After $t_f + 1$, both start decreasing but γ_t decreases as $\Delta\gamma_t \approx -\frac{0.5(a_{t_f})^2}{t^2\sigma_{sys}^2}$ while θ_t decreases as $\theta_t = \tau_t\theta_{t-1}$ so that $\Delta\theta_t = -(1-\tau_t)\theta_{t-1}$ (large decreases for large current value). The initial decrease in θ_t is usually faster than the decrease in γ_t in which case $\gamma_t - M_t\theta_t$ increases with time even after $t_f + 1$ and in such cases the change can get detected even after t_f. In practice, the assumption of PF error being negligible may not hold when tracking changed system observations, using a PF optimal for the original system, since N has been fixed for the original system's observations and with increasing rate of change or increasing total change, the PF error coefficient blows up very quickly (shown in Section V).

B. *Sufficient Conditions for Theorem 2 (ELL Error Stability)*

From Example 3.10 of [3], we can get the following sufficient conditions for R_t to be mixing:

1) π_0 has compact support
2) *and* $\psi_t(x)$ has compact support. A sufficient condition for this is that w_t has finite support, say $[-B, B]$ (e.g. truncated Gaussian noise) and $E_{x,Y_t} \triangleq h_t^{-1}([Y_t - B, Y_t + B])$ is compact. One possible sufficient condition for this is that h_t is a homeomorphism (h_t^{-1} exists and is continuous) [7].

[16]In the extreme case (for drastic changes) the PF completely loses track, i.e. the unnormalized filter kernel starts following the system model, $R_t^{c,0} \approx Q_t^0$ causing ELL to not increase above the normal value.

[17]θ_t goes as $\delta, \delta + \delta/\epsilon^2, \delta + \delta/\epsilon^2 + \delta/\epsilon^4, \delta + \delta/\epsilon^2 + \delta/\epsilon^4 + \tau\delta/\epsilon^4, \delta + \delta/\epsilon^2 + \delta/\epsilon^4 + \tau\delta/\epsilon^4 + \tau^2\delta/\epsilon^4 ...$

3) *and* given that the state transition kernel has the form $X_t = f_t(X_{t-1}) + n_t$, $f_t^{-1}(E_{x,Y_t})$ is a compact set. One possible sufficient condition for this is that f_t is a homeomorphism [7].

Now condition 2 is equivalent to assumption (iv)$''$ in Theorem 2 (posterior state space is compact). Thus if the above three conditions hold, the change lasts for a finite time and E_{x,Y_t} has a nonzero measure (implies assumption (ii) holds) then Theorem 2.2 holds.

One possible set of sufficient conditions for Theorem 2.1 are all the conditions above and the fact that p_t^c and p_t are Gaussian with $\frac{\sigma_t^c}{\sigma_t^0}$ bounded away from zero, $\mu_t^0 = 0$ and μ_t^c is finite for all t (change is an additive bias lasting for a finite time i.e. a_{t_f} is finite). These conditions follow directly by generalizing the example in Section VII-A. All of the above are very mild assumptions.

C. Generalizations

The results proved in this paper for ELL approximation errors can be generalized at two levels. First, all results of Sections IV, V and VI are true for posterior expectations of any function of the state, i.e. $[-\log p_t(x)]$ can be replaced by any other function $f(x)$. Second, $D_{Q,t}$ which measures the "rate of change" here can in general be a metric for system model error per time step (the error being introduced due to any reason). As long as the system model error lasts for a finite time, the results of this paper will apply directly.

Thus Theorems 1 and 2 can be applied to errors in approximating the posterior estimate of any function of state given past observations (MMSE estimate of the function), when using a PF with system model error. Also, Theorem 3 can be generalized to prove that the ELL approximation error, or approximation error in MMSE estimate of any function f of the state, is upper bounded by an Alpha function of the vector of system model errors per time step, $\underline{D_Q}$.

VIII. Simulation and Experimental Results

A. Example of Section VII-A

We simulated the example of Section VII-A with system noise variance, $\sigma_{sys}^2 = 0.04$, observation noise variance, $\sigma_{obs}^2 = 1$ and truncation parameter, B=10. We tested for increasing magnitudes of Δa, $\Delta a = r\sigma_{sys}$ with $r = 0$ (no change) and $r = 1, 2, 5, 10$. We show the ROC (Receiver Operating Characteristic) plots in Figure 1 for comparing performance of ELL and OL for different rates of change. A ROC for a change detection problem [8] is obtained by plotting the average detection delay against the average time between false alarms for each value of the detection threshold (varied in an appropriate range). We simulated 20 realizations and calculated the average detection delay and the average time between false alarms for different values of the detection threshold. As can be seen from the figures, the ELL (blue 'o') detects the $r = 1$ ("slow change") much faster than the OL (red '*'), ELL is slightly better than OL for $r = 2$ ("faster change") and ELL completely fails but OL detects the change almost immediately for $r = 5$ ("drastic change").

We also implemented CUSUM on ELL (green $\triangle$) and CUSUM on OL (majenta x) described in Section III-E. Once again CUSUM-ELL performs significantly better than CUSUM-OL for the slow change.

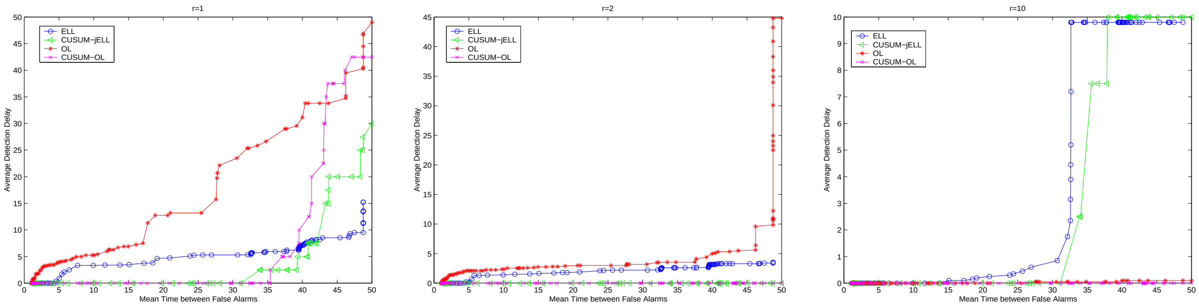
B. Abnormal Activity Detection

Now we show application of our change detection strategy to the problem of abnormal activity detection [1], in which we defined a general HMM for the normal activity. This was the practical problem that motivated the entire work described in this paper. We proposed in [1], a stochastic shape dynamical model for modeling the changing configuration of a group of moving and interacting objects. In the specific application that we experimented with, we modeled the "normal activity" of a group of passengers deplaning and moving towards the terminal in an airport (See [1] for images of the normal and abnormal activity). The shape and scaled Euclidean motion at time t constituted the state vector, i.e. $X_t = [c_t(\mu), s_t, \theta_t]$ where μ was the mean shape, c_t was the tangent coordinate of z_t (shape at time t) in the tangent space at μ, s_t was the scale and θ_t was the rotation parameter. The noisy measurements of objects' configuration formed the observation vector, Y_t. The observation model was[18]: $Y_t = h(X_t) + w_t, \quad$ where $\quad h(X_t) = z_t(c_t, \mu)s_t e^{-j\theta_t}$ and the system model was:

$$\begin{aligned}
c_t &= A_c c_{t-1} + n_t, \quad n_t \sim \mathcal{N}(0, \Sigma_{n,c,2,t}) \\
v_t &= U(\mu)c_t, \quad U(\mu) = \text{orthogonal basis}(T_\mu) \\
z_t &= (1 - v_t^* v_t)^{1/2}\mu + v_t \\
\log s_t &= \alpha_s \log s_{t-1} + (1 - \alpha_s)\mu_s + n_{s,t}, \quad n_{s,t} \sim \mathcal{N}(0, \sigma_r^2) \\
\theta_t &= \alpha_\theta \theta_{t-1} + (1 - \alpha_\theta)\mu_\theta + n_{\theta,t}, \quad n_{\theta,t} \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2)
\end{aligned} \tag{45}$$

Abnormality was defined as a change in the shape dynamics with change parameters unknown and the change being slow or drastic. We studied the problem of detecting the change in the shape introduced by one person walking away from his normal path. The speed at which the person walked away decided the rate of change. We show in Figure 2, the plots of ELL and OL to detect the abnormality for increasing rates of change (walk-away velocities). Once again, velocity=1 was a slow change which got detected by ELL much faster than OL, while for velocity=32, ELL failed and OL detected immediately.

[18]Complex number notation was used to simplify writing equations



(a) Slow Change (b) Faster Change (c) Drastic Change

Fig. 1. Simulated example: ROC curves comparing performance of ELL, OL, CUSUM on ELL, CUSUM on OL

C. *Bearings-only Target Tracking*

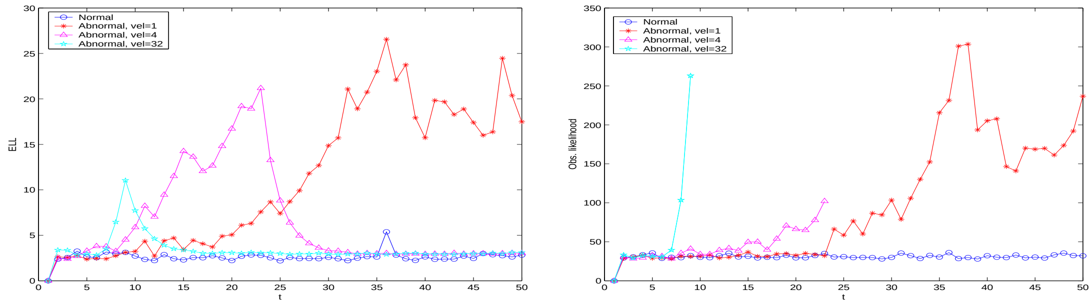
We also simulated the bearings-only target tracking example discussed in detail in [4]. In this example, the target moves on the x-y plane according to the standard second order model:

$$X_t \quad = \quad \Phi X_{t-1} + \Gamma n_t, \text{ where } \quad \Phi = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \Gamma = \begin{pmatrix} 0.5 & 0 \\ 1 & 0 \\ 0 & 0.5 \\ 0 & 1 \end{pmatrix}, \tag{46}$$

and $X_t = x_{1,t}, \dot{x}_{1,t}, x_{2,t}, \dot{x}_{1,t}$. Here $x_{1,t}$, $x_{2,t}$ denote the x and y components of the target location and $\dot{x}_{1,t}, \dot{x}_{2,t}$ denote the x and y components of the target velocity. The observation, Y_t, is a noisy measurement of the target bearing, $Y_t = \tan^{-1}(x_{2,t}/x_{1,t}) + w_t$. The system noise was zero mean i.i.d. Gaussian white noise i.e. $\Sigma_{sys} = \sigma^2_{sys} I$, $\sigma^2_{sys} = 0.001$ and the observation noise was zero mean i.i.d. Gaussian with variance $\sigma^2_{obs} = 0.005$. We attempted to detect a change in the system model where change was due to an additive bias of $[r\sigma_{sys} \ 0]^T$ added to the system noise, n_t, in (46). We show ROC plots in Figure 3 for comparing performance of ELL and OL for slow (r=1), faster (r=2) and drastic (r=20) changes. Here again, ELL and CUSUM-ELL work better than OL and CUSUM-OL for the slow change and vice versa for the drastic change.

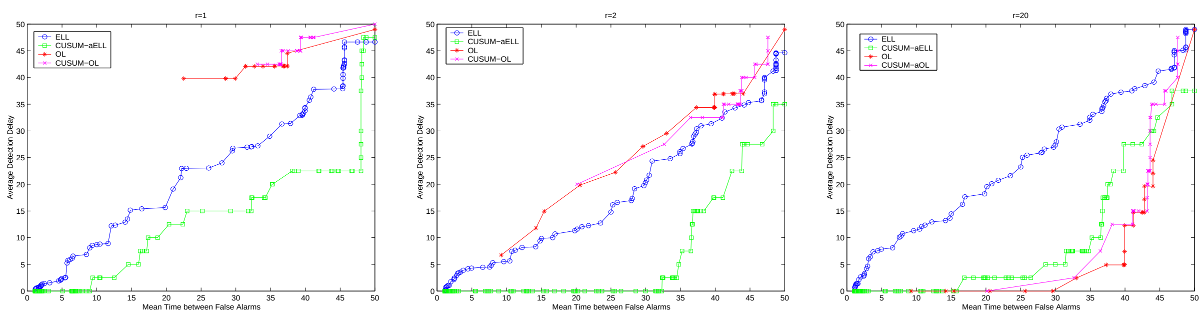
IX. Conclusions

We have studied in this paper the change detection problem in general HMMs tracked using a particle filter optimal for the unchanged system and proposed a statistic called ELL for slow change detection. We have proved in Theorems 1 and 2, the asymptotic stability and stability (under weaker assumptions) of the errors in ELL approximation. We have shown in Theorem 3 that the ELL error is upper bounded by an increasing function of



(a) ELL (b) OL

Fig. 2. Abnormal Activity Detection: Time plots of ELL and OL



(a) Slow Change (b) Faster Change (c) Drastic Change

Fig. 3. Bearings-only tracking example: ROC curves comparing ELL, OL, CUSUM on aELL, CUSUM on OL

the rate of change with all increasing derivatives. Complementary behavior of ELL and OL for change detection is discussed in Theorem 4. Simulation results on a simulated one dimensional problem, a real abnormal activity detection problem and a bearings only tracking problem have been presented. We are currently working on using ELL for neural signal processing where the goal is to detect how quickly an animal's brain responds to changes in stimuli provided to it. As part of future work, we would like to study the implications of the Alpha function bound on ELL error (Theorem 3) for improving change detection performance. Also, we would like to analyze the performance of the CUSUM algorithm on ELL and compare it with CUSUM on OL.

X. Appendix: Proofs

Proof of Theorem 1.1:

- E_{x,Y_t} being a compact and proper subset of E_t (assumption (iv)) implies that there exists $M_t < \infty$, s.t. $[-\log p_t(x)] \leq M_t$ for all $x \in E_{x,Y_t}$. Because of the uniform compactness $M^* = \sup_t M_t < \infty$. Or in other words, $[-\log p_t(x)]$ is uniformly bounded by M^* for all t.

- First consider normal observations. Since assumptions (ii) and (iii) hold and since $[-\log p_t(x)] \leq M^*$ (bounded), we can apply the Lemma 2.2 (for uniformly mixing kernels). Taking $\phi(x) = \frac{[-\log p_t(x)]}{M^*}$ [19], $\mu_t = \pi_t^0, \mu_t^N = \pi_t^{0,N}$, $\epsilon_k = \epsilon^0,\ \forall k$,we get

$$E_{Y_{1:t}}[\Xi_{pf}[|K(\pi_t^{0,N} : p_t) - K(\pi_t^0 : p_t)|]] = M^* E_{Y_{1:t}}[\Xi_{pf}[|(\pi_t^{0,N} - \pi_t^0, \frac{[-\log p_t(x)]}{M^*})|]] \leq \frac{M^*\beta^*}{\sqrt{N}}$$

  Taking $N \to \infty$, first equation of (10) follows.

- For changed observations [20],

$$|K_t^c - K_t^{c,0,N}| \quad \leq \quad |K_t^c - K_t^{c,0}| + |K_t^{c,0} - K_t^{c,0,N}| \tag{47}$$

  - Since (iii) holds, we can apply Lemma 1 with $\epsilon = \min\{\epsilon^c, \epsilon^{c,0}\}$ and $\tau = \max\{\tau^c, \tau^{c,0}\}$. We take $\phi(x) = \frac{[-\log p_t(x)]}{M^*}$, $\mu_t = \pi_t^c, \prime\mu_t = \pi_t^{c,0}$, $R_k = R_k^c, \forall t_c \leq k \leq t_f$, $R_k = R_k^{c,0}, \forall k > t_f$, and consider $t \geq t_f + 3$. Then we get

$$|K_t^c - K_t^{c,0}| \leq M^*(\tau)^{(t-t_f-3)} \sum_{k=t_c}^{t_f} (\tau)^{(t_f-k)} \delta_k \leq 2M^*(t_f - t_c + 1)(\tau)^{(-t_f-3)} \tau^t \triangleq LM^* \tau^t \tag{48}$$

    The second inequality follows because $\delta_k \leq 2$ (inequality (6)) and the fact that $\tau < 1$. For uniformly mixing kernels, ϵ and hence also τ are non-random (independent of $Y_{1:t}$) and so we can take $E_{Y_{1:t}}[.]$ in (48) and the RHS remains unchanged. Now taking $t \to \infty$, we get $\lim_{t\to\infty} E_{Y_{1:t}}[|K_t^c - K_t^{c,0}|] = 0$. i.e. that given any error $\Delta > 0$, we can choose a t_Δ s.t. $\forall t \geq t_\Delta$, $E_{Y_{1:t}}[|K_t^c - K_t^{c,0}|] \leq \Delta/2$.
  - Now fix $t = t_\Delta$, and apply Lemma 2.2 (for uniformly mixing kernels) to $|K_t^{c,0} - K_t^{c,0,N}|$ with $\mu_t = \pi_t^{c,0}, \mu_t' = \pi_t^{c,0,N}$, $R_k = R_k^0, \forall k < t_c$, $R_k = R_k^{c,0}, \forall k \geq t_c$ and $\epsilon_k = \min\{\epsilon^0, \epsilon^{c,0}\}$. Then we get: $E_{Y_{1:t}}[\Xi_{pf}[|K_{t_\Delta}^{c,0} - K_{t_\Delta}^{c,0,N}|]] \leq \frac{M^*\beta^*}{\sqrt{N}}$. Taking $N \to \infty$, we get $\lim_{N\to\infty} E_{Y_{1:t}}[\Xi_{pf}[|K_{t_\Delta}^{c,0} - K_{t_\Delta}^{c,0,N}|]] = 0$.

  Now since β^* is constant with time, the above convergence is uniform in t and so we can take $\lim_{t,N\to\infty}$ simultaneously. Thus taking $\lim_{t,N\to\infty} E_{Y_{1:t}}[\Xi_{pf}[.]]$ in (47), we get the result.

Proof of Theorem 1.2:

Since assumption (iv) (of Theorem 1.1) does not hold, $[-\log p_t(x)]$ is not bounded in this case. But we can approximate it by the increasing sequence of bounded functions $[-\log p_t^M(x)] = \min\{[-\log p_t(x)], M\}$. So we have $\lim_{M\to\infty}[-\log p_t^M(x)] = [-\log p_t(x)]$ pointwise in x.

- First consider normal observations.

$$|K_t^0 - K_t^{0,M,N}| \leq |K_t^0 - K_t^{0,M}| + |K_t^{0,M} - K_t^{0,M,N}| \tag{49}$$

[19]Note that $\phi(x) \leq 1\ \forall x \in E_{x,Y_t}$ and both posterior distributions π_t, π_t^N are zero outside E_{x,Y_t}. Hence the inner product over E is equal to the inner product taken over the set E_{x,Y_t}.

[20]For ease of notation, we denote $K(\pi_t^c : p_t)$ by K_t^c, $K(\pi_t^{c,0,N} : p_t)$ by $K_t^{c,0,N}$ and so on

– Now $K_t^{0,M}(Y_{1:t}^0) \leq K_t^0(Y_{1:t}^0)\ \forall Y_{1:t}^0$ and hence

$E_{Y_{1:t}}[|K_t^0 - K_t^{0,M}|] = |E_{Y_{1:t}}[K_t^0] - E_{Y_{1:t}}[K_t^{0,M}]| = |E_{p_t}[-\log p_t(x)] - E_{p_t}[-\log p_t^M(x)]|$

– Applying Monotone Convergence Theorem (MCT) [7](page 87), with $\mu = p_t,\ f_M(x) = [-\log p_t^M(x)]$ [21], we get $\lim_{M\to\infty} E_{Y_{1:t}}[|K_t^0 - K_t^{0,M}|] = 0$.

By assumption (iv)′, the above convergence is uniform in t. Thus given an error Δ, one can choose an M_Δ (independent of t) large enough s.t. $\forall M \geq M_\Delta$, $|K_t^{0,M} - K_t^0| < \Delta/2$.

– Now fixing $M = M_\Delta$, one can apply Theorem 1.1, with $M^* = M_\Delta$ and $p_t = p_t^{M_\Delta}$ to get that $\lim_{N\to\infty} E_{Y_{1:t}}[\Xi_{pf}[|K_t^{0,M_\Delta} - K_t^{0,M_\Delta,N}|]] = 0$, uniformly in t.

Thus taking $\lim_{M\to\infty}(\lim_{N\to\infty} E_{Y_{1:t}}[\Xi_{pf}[.]])$ in (49), we get the result.

- For changed observations,

$$|K_t^c - K_t^{c,0,M,N}| \leq |K_t^c - K_t^{c,M}| + |K_t^{c,M} - K_t^{c,0,M,N}| \tag{50}$$

– We can again apply MCT [7] with $\mu = p_t^c$ this time, to get $\lim_{M\to\infty} E_{Y_{1:t}}[|K_t^c - K_t^{c,M}|] = 0$ uniformly in t (by assumption (iv)′). Thus given an error Δ, one can choose an M_Δ, s.t. $\forall M \geq M_\Delta$, $|K_t^{c,M} - K_t^c| < \Delta/3$.

– Applying Theorem 1.1, with $M^* = M_\Delta$ and $p_t = p_t^{M_\Delta}$, we can show that $\lim_{t,N\to\infty} E_{Y_{1:t}}[\Xi_{pf}[|K_t^{c,M_\Delta} - K_t^{c,0,M_\Delta,N}|]] = 0$ [22].

Thus taking $\lim_{M\to\infty}(\lim_{t,N\to\infty} E_{Y_{1:t}}[\Xi_{pf}[.]])$ in (50), we get the result.

Proof of Theorem 1.3:

By assumption (iv)″, we have a compact posterior state space, E_{x,Y_t}, which is a proper subset of E_t, and this implies that there exists M_t s.t. $[-\log p_t(x)] < M_t,\ \forall x \in E_{x,Y_t}$. Thus the total error can be split as

$$|K_t^c - K_t^{c,0,N}| = |K_t^c - K_t^{c,0}| + |K_t^{c,0} - K_t^{c,0,N}| \tag{51}$$

Now using (48) with $M^* = M_t$, we get $|K_t^c - K_t^{c,0}| \leq LM_t\tau^t$. But by assumption (iv)″, the increase of M_t is atmost polynomial i.e. $M_t = bt^p$ for some finite p and b. It is simple to show that $M_t\tau^t$ goes to zero as t goes to infinity (apply L'Hospital's rule p times). This implies that $\lim_{t\to\infty} |K_t^c - K_t^{c,0}| = 0$.

$$\text{Also by Lemma 2.1,}\ \ \Xi_{pf}[|K_t^{c,0} - K_t^{c,0,N}|] \leq \frac{M_t\beta_t^{c,0}}{\sqrt{N}}. \tag{52}$$

Thus taking $\lim_{t\to\infty}(\lim_{N\to\infty} \Xi_{pf}[.])$ in 51, we get the result[23].

[21] Since p_t is a pdf, $\sup_x p_t(x) < \infty$. So it is easy to see that $C_t = \inf_x[-\log p_t^M(x)] > -\infty\ \forall M$, and hence we can apply MCT [7] in this case

[22] We can apply Theorem 1.1 here because M_Δ is independent of time

[23] Note that because of M_t in RHS of (52), the convergence with N is not uniform in t. We apply Lemma 2.1 to get a.s. convergence

Proof of Theorem 2.1:

The proof is similar to that of Theorem 1.2 but there are three differences. First, now the kernels $R_k^{c,0}$ are not uniformly mixing but only mixing. In this case we have for $t > t_f + 3$, $\theta_t^{c,0} = \tau_t^{c,0}\theta_{t-1}^{c,0}$. Thus $\theta_t^{c,0}$ is eventually strictly monotonically decreasing since $\tau_t^{c,0} < 1$ always. But the decrease is not exponential since $\tau_t^{c,0}$ is time varying and hence we cannot show convergence to zero of $\theta_t^{c,0}$. Secondly, now $\theta_t^{c,0}$ is a function of $Y_{1:t}$. Hence we need to take $E_{Y_{1:t}}[\theta_t^{c,0}]$. But since $\theta_t^{c,0}(Y_{1:t})$ is everywhere positive, it is trivial to show that $E_{Y_{1:t}}[\theta_t^{c,0}]$ is also eventually monotonically decreasing. The third difference here is that since $R_k^{c,0}$ is not uniformly mixing, the convergence with N is not uniform in t.

Proof of Theorem 2.2:

Now we have a bounded posterior state space at each t, i.e. $[-\log p_t(x)] < M_t, \ \forall x \in E_{x,Y_t}$. Thus the total error can be split as

$$|K_t^c - K_t^{c,0,N}| = |K_t^c - K_t^{c,0}| + |K_t^{c,0} - K_t^{c,0,N}| \tag{53}$$

Applying Lemma 1, $|K_t^c - K_t^{c,0}| \leq M_t\theta_t^{c,0}$. Applying Lemma 2.2 gives $\Xi_{pf}[|K_t^{c,0} - K_t^{c,0,N}|] \leq \frac{M_t\beta_t^{c,0}}{\sqrt{N}}$. Taking $\lim_{N\to\infty} \Xi_{pf}[.]$ in (53), we get the result.

Proof of Lemma 3:

We need to show that $f([x,y]) = \alpha_1(x, \alpha_2(y))$ is an Alpha function, given that $\alpha_1(x,z), \alpha_2(y)$ are Alpha functions of $[x,z]$ and y respectively. Consider the more general case, let

$$f([x,y]) = \sum_{j=1}^{m} \alpha_1^j(x, \alpha_2^j(y)) \tag{54}$$

and show that f is an Alpha function, given that $\alpha_1^j, \alpha_2^j, j = 1, 2, ..m$ are Alpha functions of their arguments. We prove this as follows: We show the following two facts

1) $\nabla_{x,y} f(x,y)$ (gradient of f) is an increasing function and
2) $\nabla_{x,y} f(x,y)$ can also be written as a sum of compositions of Alpha functions i.e it has the same form as f defined in (54).

Because of statement 2, the statements 1 and 2 can now be applied on ∇f to show that ∇f is an increasing function and that $\nabla\nabla f$ can also be expressed as (54). This recursive process can be continued forever to show that all derivatives of f are increasing (or that f is an Alpha function).

Proof of statement 1: Now

$$\nabla_{x,y} f(x,y) = \sum_{j=1}^{m} \begin{bmatrix} \alpha_{1x}^{j}(x, \alpha_2^j(y)) \\ \alpha_{1z}^{j}(x, \alpha_2^j(y)) \alpha_{2y}^{j}(y) \end{bmatrix} \tag{55}$$

where α_{1x}^j is partial w.r.t x and so on. It is easy to see that both the terms above are increasing functions.

Proof of statement 2: From (55), it is easy to write ∇f as a sum of compositions of Alpha functions. Setting $\tilde{\alpha}_1^j([x,y,z]) = \begin{bmatrix} \alpha_{1x}^j(x,z) \\ \alpha_{1z}^j(x,z)\alpha_{2y}^j(y) \end{bmatrix}$, $\tilde{\alpha}_2^j(y) = \alpha_2^j(y)$, we have expressed ∇f in exactly the same form as (54). We have used here the facts that derivative of an Alpha function is also an Alpha function (follows from the definition of an Alpha function) and that the product of two Alpha functions is also an Alpha function (simple to prove using an argument exactly like the one used here).

Proof of Lemma 4:

For ease of notation, denote $\sup_x \psi_{k,Y_k}(x) \triangleq S$. We first prove the following three inequalities below and then apply them to bound δ_k, ρ_k. Note that $R_{k,Y_k} = R^c_{k,Y_k^c}$ when applying Lemma 1 ("exact filtering error" bound) but $R_{k,Y_k} = R^0_{k,Y_k^c}$ when using Lemma 2 (PF error bound for incorrect model).

$$\begin{aligned} &||R^0_{Y_k^c}(\pi_{k-1}^{c,0}) - R^c_{Y_k^c}(\pi_{k-1}^{c,0})|| \leq \int_x \int_{x'} |R^0_{Y_k^c}(x,x') - R^c_{Y_k^c}(x,x')| \pi_{k-1}^{c,0}(x) dx' dx \\ &\leq \sup_x \int_{x'} |R^0_{Y_k^c}(x,x') - R^c_{Y_k^c}(x,x')| dx' \triangleq D_R(R^0_{Y_k^c}, R^c_{Y_k^c}) = D_{Q,k} \end{aligned} \tag{56}$$

$$\begin{aligned} \text{Also,}\quad |A_k - R^0_{k,Y_k^c}(\pi_{k-1}^{c,0})(E)| &= |R^c_{k,Y_k^c}(\pi_{k-1}^{c,0})(E) - R^0_{k,Y_k^c}(\pi_{k-1}^{c,0})(E)| \\ &\leq \int_{x'} | \int_x (R^0_{Y_k^c}(x,x') - R^c_{Y_k^c}(x,x')) \pi_{k-1}^{c,0}(x) dx | dx' \\ &= ||R^0_{Y_k^c}(\pi_{k-1}^{c,0}) - R^c_{Y_k^c}(\pi_{k-1}^{c,0})|| \overset{(a)}{\leq} D_{Q,k} \end{aligned} \tag{57}$$

Inequality (a) follows from of (56). Next, we lower bound $A_k = C - (C - A_k)$:

$$\begin{aligned} C - A_k = |C - A_k| &\leq ||R^c_{k,Y_k^c}(\pi_{k-1}^c - \pi_{k-1}^{c,0})|| \overset{(b)}{\leq} \frac{\lambda^c_{k,Y_k^c}(E) ||\pi_{k-1}^c - \pi_{k-1}^{c,0}||}{\epsilon_k^c} \triangleq \frac{\tilde{D}_{k-1}}{\epsilon_k^c} \\ \text{Thus,}\quad A_k &\geq C - \frac{\tilde{D}_{k-1}}{\epsilon_k^c} \end{aligned} \tag{58}$$

(b) follows from Lemma 3.5 of [3] and mixing property of R_k.

Now we use the above inequalities to bound δ_k:

$$\begin{aligned}\delta_k = \sup_{\phi:||\phi||_\infty \leq 1} |(\pi_k^{c,0} - \bar{R}^c_{Y_k^c}(\pi_{k-1}^{c,0}), \phi)| &\leq ||\pi_k^{c,0} - \bar{R}^c_k \pi_{k-1}^{c,0}|| = ||\bar{R}^0_{Y_k^c}(\pi_{k-1}^{c,0}) - \bar{R}^c_{Y_k^c}(\pi_{k-1}^{c,0})|| \\ &\overset{(c)}{\leq} \frac{||R^0_{Y_k^c}(\pi_{k-1}^{c,0}) - R^c_{Y_k^c}(\pi_{k-1}^{c,0})|| + |A_k - R^0_{k,Y_k^c}(\pi_{k-1}^{c,0})(E)|}{A_k} \\ &\overset{(d)}{\leq} \frac{2D_{Q,k}}{A_k} \overset{(e)}{\leq} \frac{2D_{Q,k}}{C - \frac{\tilde{D}_{k-1}}{\epsilon_k^c}} \end{aligned} \tag{59}$$

Inequality (c) is an application of inequality (6) of [3] (given in (32)), (d) follows by combining (56) and (57) and (e) follows from (58). Now consider ρ_k:

$$\rho_k \overset{(f)}{\leq} \frac{S}{{\epsilon_k^{c,0}}^2 R^0_{k,Y_k^c}(\pi_{k-1}^{c,0})(E)} \overset{(g)}{\leq} \frac{S}{{\epsilon_k^{c,0}}^2 (A_k - D_{Q,k})} \overset{(h)}{\leq} \frac{S}{{\epsilon_k^{c,0}}^2 (C - \frac{\tilde{D}_{k-1}}{\epsilon_k^c} - D_{Q,k})}$$

Inequality (f) follows from Remark 5.10 of [3] (given in (31)), (g) follows from (57) and assumption (18); (h) follows from (58) and assumption (18).

Also note that it is easy to see that $f(z) = \frac{a}{(b-cz)}$ and also $f(z) = az$ is an Alpha function. Thus the bound on δ_k is an Alpha function of $\tilde{D}_{k-1}$ and $D_{Q,k}$. The bound on ρ_k is an Alpha function of $\frac{1}{\epsilon}$ since $f(z) = z^2$ is an Alpha function; it is an Alpha function of $D_{Q,k}$ and $\tilde{D}_{k-1}$ since $f(z) = a/(b - cz)$ is an Alpha function.

References

[1] N. Vaswani, A. RoyChowdhury, and R. Chellappa, "Activity recognition using the dynamics of the configuration of interacting objects," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Madison, WI, June 2003.

[2] N. Vaswani, *Change Detection in Stochastic Shape Dynamical Models with Applications in Activity Modeling and Abnormality Detection*, Ph.D. Thesis, ECE Dept, University of Maryland at College Park, August 2004.

[3] LeGland F. and Oudjane N., "Stability and Uniform Approximation of Nonlinear Filters using the Hilbert Metric, and Application to Particle Filters," *Technical report, RR-4215, INRIA*, 2002.

[4] N.J. Gordon, D.J. Salmond, and A.F.M. Smith, "Novel approach to nonlinear/nongaussian bayesian state estimation," *IEE Proceedings-F (Radar and Signal Processing)*, pp. 140(2):107–113, 1993.

[5] D.F. Kerridge, "Inaccuracy and inference," *J. Royal Statist. Society, Ser. B*, vol. 23 1961.

[6] Rudolf Kulhavy, "A geometric approach to statistical estimation," in *IEEE Conference on Decision and Control (CDC)*, Dec. 1995.

[7] H.L. Royden, *Real Analysis*, Prentice Hall, 1995.

[8] M. Basseville and I Nikiforov, *Detection of Abrupt Changes: Theory and Application*, Prentice Hall, 1993.

[9] B. Azimi-Sadjadi and P.S. Krishnaprasad, "Change detection for nonlinear systems: A particle filtering approach," in *American Control Conference*, 2002.

[10] A. Doucet, N. deFreitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice*, Springer, 2001.

[11] Y. Bar-Shalom and T. E. Fortmann, *Tracking and Data Association*, Academic Press, 1988.

[12] C. Andrieu, A. Doucet, S.S. Singh, and V.B. Tadic, "Particle methods for change detection, system identification, and control," *Proceedings of the IEEE*, vol. 93, pp. 423– 438, March 2004.

[13] D. Ocone and E. Pardoux, "Asymptotic stability of the optimal filter with respect to its initial condition," *SIAM Journal of Control and Optimization*, pp. 226–243, 1996.

[14] Rami Atar and Ofer Zeitouni, "Lyapunov Exponents for Finite State Nonlinear Filtering," *SIAM Journal on Control and Optimization*, vol. 35, no. 1, pp. 36–55, 1997.

[15] F. LeGland and L. Mevel, "Exponential forgetting and geometric ergodicity in hidden markov models," *Mathematics of Control, Signals and Systems*, pp. 63–93, 2000.

[16] R. Atar, "Exponential stability for nonlinear filtering of diffusion processes in a noncompact domain," *Ann. Probab.*, pp. 1552–1574, 1998.

[17] A. Budhiraja and D. Ocone, "Exponential stability of discrete time filters for bounded observation noise," *System and Control Letters*, pp. 185–193, 1997.

[18] P. DelMoral, "Non-linear filtering: Interacting particle solution," *Markov Processes and Related Fields*, pp. 555–580, 1996.

[19] N. Vaswani, "Change detection in partially observed nonlinear dynamic systems with unknown change parameters," in *American Control Conference (ACC)*, 2004.

[20] G. Casella and R. Berger, *Statistical Inference*, Duxbury Thomson Learning, second edition, 2002.

[21] T. Cover and J. Thomas, *Elements of Information Theory*, Wiley Series, 1991.

[22] N. Vaswani, "Bound on errors in particle filtering with incorrect model assumptions and its implication for change detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2004.

[23] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, Inc., 1991.